Investigating the Non-Stationary Bandit Problem

Lancaster University

Dimitri Zografos

August 28, 2020

Contents

1 Introduction **5**

2 Definition of the Bandit Problem and Existing Strategies **7**
2.1 Definition of the bandit problem 7
2.1.1 Stochastic Bandit Problem 8
2.1.2 Adversarial Bandit Problem 9
2.2 Existing Strategies for Stationary Stochastic Bandits 10
2.2.1 Construction of Upper Confidence Bound Strategies 10
2.2.2 UCB1 and Regret Analysis 11
2.2.3 Lower bound on the Regret 13
2.2.4 Alternative type of policies 16
2.2.5 KL-UCB 18
2.3 Strategies for Adversarial Bandits 21

3 Stochastics Bandits: Non-Stationary Case **25**
3.1 Formulation of the Non-Stationary case 26
3.2 Discounted UCB and Sliding-Window UCB 26
3.3 Lower Bound on the Regret of the Piecewise-Stationary Setting 32
3.4 Monitored-UCB 35
3.4.1 Proof of Theorem 3.4.1 42

4 Experimental Results **45**
4.1 Simulations 45
4.2 Fixed number of change-points generated at random 49
4.3 Extension: Random number of change-points 56

5 Conclusions **61**

Chapter 1

Introduction

The aim of this work is to describe the Non-Stationary Bandit Problem with its properties and investigate how policies can perform in such a setting. To do so we are going to show first some policies for both the stochastic and the adversarial one providing a theoretical analysis about the construction of said policies. These policies will be used to construct the strategies for the Non-Stationary setting.

More specifically, in Chapter 2 we are going to define what the multi-armed bandit problem is and the main differences between the stochastic setting and the adversarial one. Following this, we are going to picture the construction of the Upper Confidence Bound (UCB) type of policies with a description of the main ones like UCB1, (α, ψ)-UCB and KL-UCB (Auer et al. [2002a], Bubeck et al. [2012], Garivier and Cappé [2011]). To give a context to the performances of the algorithms we are also going to provide a lower bound on the possible expected regret incurred during the game by any policy in the stochastic stationary setting. We will describe the ϵ-Greedy to highlight the different approach yielded by the two different strategies and how they apply in a different way the concepts of exploration and exploitation. This chapter is concluded by describing the design of policies for the adversarial setting and by defining EXP3 and showing its regret analysis.

Chapter 3 is going to be the core of the work and will deal with the Non-Stationary setting. After a preliminary definition of notation we will start by understanding why the policies for the standard setting would not work in this scenario. We will then move to describe two of the main policies designed for this setting: D-UCB and SW-UCB, presented in Garivier and Moulines [2011]. After analysing similarities and differences of these two UCB based methods we will provide a regret analysis for the former. The focus will then move to analyse the lower-bound on the expected regret obtainable by any policy in this setting. We will conclude this chapter by showing the M-UCB strategy, which was presented in Cao et al. [2019] and uses an alternative approach based on a combination of the regular UCB and a change-point detector.

In Chapter 4 we will start by observing how the M-UCB algorithm reacts if we stress the assumption made or if we make minor changes to the algorithm itself. More specifically we will lead simulations on how the policy would perform if we only restart the arms where the changes have been spotted and how the M-UCB algorithm would perform if the change-points were generated at random. After pointing out the symptoms that these modifications cause we will try to look at the latter change with a more theoretical approach. To do so we will introduce the concepts of uniform spacings and the probability distribution of the length of the shortest subsegment of a segment broken in m parts which were discussed first in [Pyke, 1965] and then deepen in Pyke et al. [1972]. We will then discuss the M-UCB algorithm in this different setting considering the probabilistic nature that the problem assume under this new constraint. After the regret analysis

of M-UCB in this setting we will observe that if the decision maker plays the game for a fraction of the time segment only, the number of changes occurring becomes random. We will end the work by introducing a possible approach for that based on a setting the number of changepoints in the system to a Bayesian mixture of the possible numbers of changepoints. The parameter of each component q of the mixture at each time t will be given by the updated probability that the segments observed by time t have been generated by a system with q changes total.

Chapter 2

Definition of the Bandit Problem and Existing Strategies

In this chapter we are going to provide the definitions of what a bandit problem is and of its main features. After defining the problem and giving some general notions like the ones of *regret* and *strategy* we will distinguish between the stochastic setting and the adversarial one. For this part, we will mainly reference to Bubeck et al. [2012] and Auer et al. [2002a].

2.1 Definition of the bandit problem

A *bandit problem* is a sequential decision problem with limited information available. If the choices available at each time step are multiple then the bandit problem is called *multi-armed*.
The name of the problem derives from slot-machines being called bandits in American slang and it is useful to give an immediate insight of it. If the decision maker can only play one slot machine then the problem is a "one-armed" one as the player can only either choose between playing that slot machine or not playing whereas, if he can choose between multiple slot machines the problem is "multi-armed".
The goal of the decision maker is to maximize his payoff through the sequential choices he makes which leads to one of the most important components of the bandit problem. In fact, the player has to find the optimal compromise between the exploration of all arms and the exploitation of the arm he perceives as the best to maximize his payoff. More precisely, if he does not spend a sufficient amount of time exploring all arms he might not have enough information about all the choices he can make whereas, if he spends too much time exploring and not enough exploiting the arm he perceives as optimal he would not benefit enough of the optimal arm itself.
Bandit problems have a wide pool of applications: from ad placement to clinical trials, from source routing to live-strategic videogames. All these different problems require different developments of the problem which implies that the bandit problem can be formalized in a lot of different ways. The first distinction we make considers two of the most used natures of payoff processes: the stochastic and the adversarial one.
Before moving to analyze the differences between these two different settings we define the notion of *regret*. By regret we indicate the number of times the forecaster makes a suboptimal choice. Therefore, if our goal is to maximize the payoff of the decision maker, it is also to minimize the amount of time a suboptimal decision is taken and thus, the regret.
Formally, given a number of arms K and a sequence $X_{k,t}$ where $k \leq K$ and $t = 1, 2, ...$ representing the unknown pay-offs associated with arm k sampled at time t and calling the choice of the

player at each time as π_t with reward $X_{\pi_t,t}$, we define the regret after T steps as

$$R(T) = \max_{k=1,\dots,K} \sum_{t=1}^{T} X_{k,t} - \sum_{t=1}^{T} X_{\pi_t,t}. \tag{2.1}$$

If the decision taken by the decision maker at each step follows an algorithmical logical scheme, we call π_t a *policy*.

Different assumptions on the nature of the rewards and decisions yield different definitions of regret as well. Next, we are going to define two of the main settings: the stochastic and the adversarial one. Even if the main focus of the dissertation is going to be on stochastic bandits especially when dealing with the non stationary scenario, we will also describe the most famous algorithm for the adversarial setting.

2.1.1 Stochastic Bandit Problem

The first setting we are going to define is the *stochastic* one. This setting relies on the assumptions that both the sequence of rewards $(X_{k,t})_t$ and the player choices π_t are random variables. These assumptions allow us to define functions of the regret based on the expectation of these random variables like the *expected regret*

$$\mathbb{E}[R(T)] = \mathbb{E}\left[\max_{k=1,\dots,K} \sum_{t=1}^{T} X_{k,t} - \sum_{t=1}^{T} X_{\pi_t,t}\right] \tag{2.2}$$

ant the *pseudo-regret*

$$\overline{R}(T) = \max_{k=1,\dots,K} \mathbb{E}\left[\sum_{t=1}^{n} X_{k,t} - \sum_{t=1}^{T} X_{\pi_t,t}\right]. \tag{2.3}$$

We can notice that the definition of the expected regret is stronger as it evaluates the expectation of the regret calculated with respect to the optimal arm at each step whereas the pseudo-regret is calculated with respect to the arm which maximizes that difference between playing that arm n times and the sequence of decision performed by the forecaster. As a formal implication, $\overline{R}(T) \leq \mathbb{E}[R(T)]$.
In the general stochastic setting we do not make any assumption about the mean reward of an arm $k \leq K$ remaining the same over time and for this reason we define $\mu_{k,t} = \mathbb{E}[X_{k,t}]$.
To get started, we are going to assume that the pay-offs $X_{k,t},\ t \in \{1, ..., T\}$ are independent and identically distributed for every $k \leq K$. The non-stationary case is going to be studied further in the work. If we assume the mean of the random variable of the rewards to be stationary over time for each arm k then we can define the mean of each arm as $\mu_{k,t} = \mathbb{E}[X_{k,t}]$.
We can therefore introduce the notions of *optimal mean* μ^* as

$$\mu^* = \max_{k=1,\dots,K} \mu_k \tag{2.4}$$

and of *optimal decision* as

$$\pi^* \in \operatorname*{argmax}_{k=1,\dots,K} \mu_k. \tag{2.5}$$

According to this assumption we can also rewrite the definition we gave of *pseudo-regret*,

$$\overline{R}(T) = T\mu^* - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{\pi_t}\right]. \tag{2.6}$$

This simplification of the definition of pseudo-regret says that the regret that a decision maker is going to incur depends on the number of times he did not play the optimal arm π^*. Therefore, the pseudo-regret in the stationary stochastic setting only depends on the policy π_t used by the forecaster. Before moving to the definition of the adversarial problem we give a formal statement of the stochastic one.

Definition 2.1.1 (Stationary Stochastic Bandit Problem)**.** *Let K be the number of different playable arms and, if available, let T be the number of rounds the decision-maker has to play. Assume that the rewards coming from different arms follow unknown distributions $P^{(1)}, ..., P^{(K)}$ with unknown means $\mu_1, ..., \mu_K$. At each time step $t = 1, ..., T$, the decision maker*

1) *chooses one arm to play $\pi_t \in \{1, ..., K\}$;*

2) *plays arm π_t and receives a reward $X_{\pi_t,t} \sim P^{(\pi_t)}$ from the environment independently of what happened in the previous rounds. The other rewards are not going to observed*

2.1.2 Adversarial Bandit Problem

While the Stochastic setting was based, as the name suggests, on probabilistic assumptions made on the rewards, the Adversarial setting does not rely on any of these. Instead, imagine that we are playing a game where the environment, that can also be considered as an opponent, chooses at each time step t what the rewards $x_{k,t}$ are going to be for each $k \in \{1, ..., K\}$. We decided to change the notation of the rewards to clench that the rewards are not random variables. Even if the environment might want to set the rewards in a tricky way to make the decision-maker lose, it can not put all the rewards to zero as no people would play the game and it would be counter-productive.
It is clear then that the adversarial type of problem presents different difficulties. To make an example, in the adversarial case we also have to question ourselves whether the environment is going to adapt to the choices made by the player or not. If the opponent is going to adapt to the decisions of the forecaster than we define the environment as *nonoblivious*, whereas if the environment is going to set the rewards independently of the player strategy we call it *oblivious*. The distinction between those two scenarios is very important as one is possibly much harder to deal with than the other. Specifically, dealing with a nonoblivious environment might include a lot of dependencies to be understood and studied as the reward x_{π_t} is not a function of all the decisions taken in the previous rounds, which is $x_t = x_t(\pi_1, ..., \pi_{t-1})$. If we are dealing with a nonoblivious environment then it might be very hard to upper-bound the *expected regret* because of all the possible combinations of rewards and also because the strategy of the opponent is going to depend on player's one. Is therefore preferable to use the pseudo-regret

$$\overline{R}(T) = \max_{k=1,...,K} \mathbb{E}\left[\sum_{t=1}^{n} x_{k,t} - \sum_{t=1}^{T} x_{\pi_t,t}\right]. \tag{2.7}$$

Definition 2.1.2 (Adversarial Bandit Problem)**.** *Let K be the number of different playable arms and, if available, let T be the number of rounds the decision-maker has to play. At each time step $t = 1, ..., T$,*

1) *the decision maker chooses one arm to play $\pi_t \in \{1, ..., K\}$;*

2) *in the same moment the opponent sets the rewards $x_{1,t}, ..., x_{K,t}$;*

3) *the player plays arm π_t and observes and receives a reward $x_{\pi_t,t}$ from the rewards set previously by the environment. The other rewards are not going to be observed.*

2.2 Existing Strategies for Stationary Stochastic Bandits

In this section we dwell on the stationary stochastic setting.
We will briefly recall some of the definitions given in the previous section and introduce some new ones. We will then move on to present some fundamental results and then some of the most important policies for this type of problem.
Earlier we introduced some notation for the optimal mean and decision as

$$\mu^* = \max_{k=1,\dots,K} \mu_k \tag{2.8}$$

and of *optimal decision* as

$$\pi^* \in \max_{k=1,\dots,K} \mu_k. \tag{2.9}$$

Let us define the gap that the mean of each arm k has from the optimal one as

$$\Delta_k = \mu^* - \mu_k \tag{2.10}$$

and the number of times arm k has been played in t rounds as

$$N_k(t) = \sum_{s=1}^{t} \mathbb{1}\{\pi_s = k\}. \tag{2.11}$$

Then, if we consider the regret for the stationary stochastic case, we can write a simplification using the notation just introduced

$$\overline{R}(T) = T\mu^* - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{\pi_t}\right] = \sum_{k=1}^{K} \mathbb{E}\left[N_k(T)\right] \mu^* - \sum_{k=1}^{K} \mathbb{E}\left[N_k(T)\right] \mu_k \tag{2.12}$$

$$= \sum_{k=1}^{K} \Delta_k \mathbb{E}\left[N_k(T)\right]. \tag{2.13}$$

Given the sequence of rewards obtained $(X_{\pi_t,t})_t$ we can break it in K sequences $(Z_{\pi_t,N_{\pi_t}(t)})_{N_{\pi_t}(t)}$, one per arm, in order to track separately the different payoffs drawn from the arms.
We also introduce the notation $\hat{\mu}_k(t)$, which represents the empirical mean of rewards of arm k after pulling it t times. As we are considering rewards which are i.i.d. we have that $\hat{\mu}_k(t) = \frac{1}{t}\sum_{s=1}^{t} Z_{k,s}$ and $\hat{x}_k(t) = \frac{1}{N_k(t)}\sum_{s=1}^{t} X_{\pi_s,s}\mathbb{1}\{\pi_s = k\}$.
As we mentioned earlier, the decision-maker facing a bandit problem has to find a compromise between the *exploration* and *exploitation*, which means a compromise between playing the arm that he perceives as best and playing the other arms to gain information about other arms to make sure he is playing the one with the highest mean.

2.2.1 Construction of Upper Confidence Bound Strategies

The principle used on which are based most policies/strategies is the so-called *optimism in face of uncertainty*. This principle states that if the decision-maker has to choose between different environments after having gained enough information, he will have to choose the most favourable one. This simple principle yields almost optimal policies in the stochastic setting.
Specifically, using this principle, we want to find for every option we can choose at every time step, a dynamic upper-bound on each of these and choose the option with the highest bound.

To do so, we make some assumptions on the distribution of the rewards X. We assume that there exists a convex function ψ such that

$$\begin{cases} \ln \mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}[X])\right)\right] \leq \psi(\lambda) \\ \ln \mathbb{E}\left[\exp\left(\lambda(\mathbb{E}[X] - X)\right)\right] \leq \psi(\lambda) \end{cases} \tag{2.14}$$

Specifically, if we consider $X \in [0, 1]$ and $\psi(\lambda) = \frac{\lambda^2}{8}$ we have that Equation (2.14) coincides with the Hoeffding's lemma.
We want to use this equation and this assumption to build an upper-bound estimate on the mean of each arm at a given confidence level and then choose the arm which has the highest one.
We define the Legendre-Fenchel transform of the above-mentioned function ψ as

$$\psi^*(x) = \sup_{\lambda \in R}(\lambda x - \psi(\lambda)). \tag{2.15}$$

Using the Legendre-Fenchel transform, Markov's Inequality and Equation (14) we can state

$$\mathbb{P}(\mu_k - \hat{\mu}_k(t) > \epsilon) \leq \exp\left(-t\, \psi^*(\epsilon)\right). \tag{2.16}$$

This means that with probability not smaller than $1 - \delta$

$$\hat{\mu}_k(t) + (\psi^*)^{-1}\left(\frac{1}{t}\ln\frac{1}{\delta}\right) > \mu_k. \tag{2.17}$$

Thus, this leads to the criterion used for the Upper Confidence Bound (UCB) type of policies.
We will now introduce a general UCB policy and then analyse a specific one. The general strategy is called $(\alpha, \psi) -$ UCB.

Algorithm 1 $(\alpha, \psi) -$ UCB

Require: T, K, ψ and α.
Ensure: Play each arm once.

1: **for** $t = 1, 2, ..., T$ **do**
2: Let, $\forall k \leq K$,

$$UCB_k \leftarrow \hat{x}_k(t) + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)}\right)$$

3: Set $\pi_t \leftarrow \underset{k \leq K}{\text{argmax}}\, UCB_k$

2.2.2 UCB1 and Regret Analysis

A more specific version of the $(\alpha, \psi) -$ UCB is the so called UCB1 introduced in Auer et al. [2002a].
For Algorithm 2 we will also provide an upper-bound on its expected regret.

Algorithm 2 UCB1

Require: T, K.

Ensure: Play each arm once.

1: **for** $t = 1, 2, ..., T$ **do**

2: Let, $\forall k \leq K$,

$$UCB_k \leftarrow \hat{x}_k(t) + \sqrt{\frac{2 \ln t}{N_k(t-1)}}$$

3: Set $\pi_t \leftarrow \underset{k \leq K}{\text{argmax}}\, UCB_k$

Theorem 2.2.1 (Auer et al. [2002a]). *If policy UCB1 is played on K arms with reward distributions $P^{(1)}, ..., P^{(K)}$ with support in $[0, 1]$, then its expected regret after T rounds is at most*

$$\left[8 \sum_{k:\Delta_k > 0} \left(\frac{\ln T}{\Delta_k} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{k:\Delta_k > 0} \Delta_k \right). \tag{2.18}$$

Proof. Denote $c_{t,s} = \sqrt{(2 \ln t)/s}$. In order to establish the upper bound on the regret we first upper-bound the number of times each arm can be played while being suboptimal. Specifically, we bound the indicator function of $\pi_t = k$.

$$N_k(T) = 1 + \sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k\} \tag{2.19}$$

$$\leq l + \sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k, N_k(t) \geq l\} \tag{2.20}$$

$$\leq l + \sum_{t=K+1}^{T} \mathbb{1}\left\{ \hat{x}^*_{t-1} + c_{t-1,N^*(t-1)} \leq \hat{x}_{k,t-1} + c_{t-1,N_k(t-1)}, N_k(t) \geq l \right\} \tag{2.21}$$

$$\leq l + \sum_{t=K+1}^{T} \mathbb{1}\left\{ \min_{0<s<t} \hat{x}^*_{s-1} + c_{t-1,N^*_{s-1}} \leq \max_{0<s_k<t} \hat{x}_{k,s_k-1} + c_{t-1,N_k(s_k-1)}, N_k(t) \geq l \right\} \tag{2.22}$$

$$\leq l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_k=l}^{t-1} \left\{ \hat{x}^*_s + c_{t,s} \leq \hat{x}_{k,s_k} + c_{t,s_k} \right\}. \tag{2.23}$$

The first inequality follows from the sum of $l - 1$ indicator functions being upper-bounded by $l - 1$. The second is given by the definition of π_t, as, if $\pi_t = k$, then k is the arm with the highest confidence bound at time t; this already shows that the number of times an arm is perceived as the optimal one can be upper-bounded by the number of times an arm is perceived as better than a specific one. The third bound can be thought as upper-bounding the number of times the suboptimal arm has been better than the optimal one at a specific t with the number of times the maximum value of the suboptimal arm between time 0 and t has been better than the minimum of the optimal one in the same interval.

The event written in the brackets of equation (2.23) can happen only if at least one of the following events is verified:

$$\hat{x}_{k,s_k} \geq \mu_k - c_{t,s_k}; \tag{2.24}$$

$$\hat{x}^*_s \leq \mu^* - c_{t,s}; \tag{2.25}$$

$$\mu^* \leq \mu_k + 2c_{t,s_k}. \tag{2.26}$$

One of these events happening would mean that either the suboptimal arm has been overestimated (inequality (2.24)), either the optimal one has been underestimated (inequality (2.25)) or the means

of the two arms are not far apart enough (inequality (2.26)). The probabilities of events (2.24) and (2.25) can be bounded using Chernoff-Hoeffding bound

$$\mathbb{P}(\hat{x}_s^* \leq \mu^* - c_{t,s}) \leq \exp\left(-2(c_{t,s})^2 \cdot s\right) \tag{2.27}$$

$$\leq \exp\left(-4\frac{\ln t}{s}s\right) \tag{2.28}$$

$$= \exp\left(-4\ln t\right) \tag{2.29}$$

$$= t^{-4}. \tag{2.30}$$

For the third event we can set l to be a value such that if the player has played arm k a number of times equal to l, the exploration bound $c_{t,l}$ is tight enough to make the means of the two arms sufficiently far apart from each other and thus make the event impossible to happen.
Specifically, let $l = \left\lceil \frac{8\ln n}{\Delta_k^2} \right\rceil$, then

$$\mu^* - \mu_k - 2c_{t,s_k} = \mu^* - \mu_k - 2\sqrt{(2\ln t)/s_k} \geq \mu^* - \mu_k - \Delta_k = 0 \tag{2.31}$$

for every $s_k > \left\lceil \frac{8\ln T}{\Delta_k^2} \right\rceil$ which makes event (2.26) impossible.

We now pass to the expectation of $N_k(T)$. Let $\varphi = \lceil 8\ln T/\Delta_k^2 \rceil$.

$$\mathbb{E}[N_k(T)] \leq \left\lceil \frac{8\ln n}{\Delta_k^2} \right\rceil \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_k=\varphi}^{t-1} \left(\mathbb{P}(\hat{x}_s^* \leq \mu^* - c_{t,s}) + \mathbb{P}(\hat{x}_{k,s} \geq \mu_k - c_{t,s_k})\right) \tag{2.32}$$

$$\leq \varphi \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_k=\varphi}^{t-1} 2t^{-4} \tag{2.33}$$

$$\leq \varphi + 1 + \frac{\pi^2}{3} \tag{2.34}$$

which concludes the proof of the statement. □

Theorem 2.2.1 basically states that if we perform $UCB1$ policy in a stochastic stationary setting, the average number of mistakes the decision maker is going to do is not bigger than $O(lnT)$.
One might wonder if it is possible to have a policy that achieves a lower expected regret than $UCB1$ in the same setting and, if the answer is yes, what is the maximum improvement we can in obtain. In order to reply to those questions we might want to try to obtain a lower bound on the expected regret that we can possibly have.

2.2.3 Lower bound on the Regret

The lower bound result we are going to analyze in this section holds in a stochastic stationary setting with Bernoulli rewards distributions. For any couple $p, q \in [0, 1]$ we define the Kullback-Leibler divergence between two Bernoulli probability distributions P and Q with parameters p and q as

$$D(P||Q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

It can also be denoted as $D(p||q)$.
The following lower bound theorem is given with respect to the pseudo-regret.

Theorem 2.2.2 ([Bubeck et al., 2012]). *Consider a policy π so that $\mathbb{E}[N_k(T)] = o(T^\alpha)$ for any set of Bernoulli rewards distributions, any arm k with $\Delta_k > 0$ and $\alpha > 0$. Then, for any set of Bernoulli reward distributions, we can state*

$$\underset{T\to\infty}{lim}\frac{\overline{R}_T}{\ln T} \geq \sum_{k:\Delta_k>0}\frac{\Delta_k}{D(P^{(k)}||P^*)}$$

Before getting started with the proof, we recall the maximal version of the strong law of large numbers as it will be used for the proof.

Proposition 2.2.1 (Strong Law of Large Numbers for Maximals)**.** *For any sequence of independent real valued random variables (X_t) with mean $\mu > 0$ we have that*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{a.s.} \mu \quad \Rightarrow \quad \lim_{T\to\infty} \frac{1}{T} \max_{s=1,\dots,T} \sum_{t=1}^{s} X_t \xrightarrow{a.s.} \mu \tag{2.35}$$

Proof of Theorem 2.2.2. For ease of explanation we will consider a problem with two arms.
Suppose $\mu_1 > \mu_2$ and that therefore the first arm is the optimal one. We can define a continuous map $x \to D(\mu_2||x)$ and consider a Bernoulli distribution with mean $\mu_2' > \mu_1$ so that, for some $\epsilon > 0$,

$$D(\mu_2||\mu_2') \leq (1+\epsilon)D(\mu_2||\mu_1). \tag{2.36}$$

The goal we have is to compare the two different bandit problems. If we manage to show that with high probability the algorithm cannot distinguish between the two problems than we can also say that, as we know by hypothesis that in one case the algorithm yields good performances, it will perform good results on the modified problem as well.
Let us consider the sequence $Z_{2,1}, ..., Z_{2,s}$ of the first $s \in \{1, ..., T\}$ pay-offs coming from arm 2. Then we can write an empirical estimate of $D(\mu_2||\mu_2')$ as

$$\hat{D}_s(\mu_2||\mu_2') = \sum_{t=1}^{s} \frac{\mu_2 Z_{2,t} + (1-\mu_2)(1-Z_{2,t})}{\mu_2' Z_{2,t} + (1-\mu_2')(1-Z_{2,t})}. \tag{2.37}$$

At the end of the game the empirical estimate of the K-L divergence between two distributions with means μ_2 and μ_2' is given by $\hat{D}_{N_2(T)}(\mu_2||\mu_2')$. We also define a probability measure over the elements of the σ-algebra generated by the $Z_{2,1}, ..., Z_{2,T}$. For any event A in the above mentioned sigma algebra, its probability is determined by

$$\mathbb{P}'(A) = \mathbb{E}\left[\mathbb{1}_A \exp(-\hat{D}_{N_2(T)}(\mu_2||\mu_2'))\right] \tag{2.38}$$

We now introduce an event which will help link the behaviour of the player in the two different bandits as

$$C_T = \left\{ N_2(T) \leq \frac{1-\epsilon}{\hat{D}_{N_2(T)}(\mu_2||\mu_2')} \ln T \text{ and } \hat{D}_{N_2(T)}(\mu_2||\mu_2') \leq \left(1 - \frac{\epsilon}{2}\right) \ln T \right\}. \tag{2.39}$$

This event basically wants the number of times arm 2 has been played to be bounded by a quantity which is directly proportional to the logarithm of T and is inversely proportional to the estimate of the K-L divergence between the old arm 2 and the modified one. This is correct because for a UCB type of policy the number of times a suboptimal arm is played is on average $\ln T$. However, as we want to connect the old game with the new one, we must rescale this quantity according to how different the two games are. Another condition given by the event is that the estimate of the K-L divergence is smaller than a quantity dependent on $\ln T$ again.
Because of (2.38) and (2.39) we can write

$$\mathbb{P}'(C_T) = \mathbb{E}[\mathbb{1}_{C_T} \exp(-\hat{D}_{N_2(T)}(\mu_2||\mu_2'))] \geq \exp\left(-\left(1 - \frac{\epsilon}{2}\right) \ln T\right) \mathbb{P}(C_T) \tag{2.40}$$

where $\mathbb{P}'$ and $\mathbb{E}'$ are used to indicate that we are integrating with respect to the modified bandit problem.
For ease of presentation, we introduce the shorthand notation

$$f_T = \frac{1-\epsilon}{\hat{D}_{N_2(T)}(\mu_2||\mu_2'))} \ln T. \tag{2.41}$$

Equation (2.40) and Markov's Inequality allow us to say that

$$\mathbb{P}(C_T) \leq T^{(1-\frac{\epsilon}{2})}\mathbb{P}'(C_T) \leq T^{(1-\frac{\epsilon}{2})}\mathbb{P}'(N_2(T) < f_T) \tag{2.42}$$

$$\leq T^{(1-\frac{\epsilon}{2})}\frac{\mathbb{E}'[T - N_2(T)]}{T - f_T} \tag{2.43}$$

Recall that for every suboptimal arm we have that, by hypothesis, $\mathbb{E}[N_k(T)] = o(T^\alpha)$ for any $\alpha > 0$.
Therefore, we can state

$$\mathbb{P}(C_T) \leq T^{(1-\frac{\epsilon}{2})}\frac{\mathbb{E}'[T - N_2(T)]}{T - f_T} \tag{2.44}$$

We can notice that

$$\mathbb{P}(C_T) \geq \mathbb{P}\left(N_2(T) \leq f_T \text{ and } \max_{s\leq f_T} \hat{D}_s(\mu_2||\mu_2')) \leq \left(1 - \frac{\epsilon}{2}\right)\ln T\right) \tag{2.45}$$

$$= \mathbb{P}\left(N_2(T) \leq f_T \text{ and } \frac{D(\mu_2||\mu_2')}{(1-\epsilon)\ln T} \times \max_{s\leq f_T} \hat{D}_s(\mu_2||\mu_2')) \leq \frac{1-\frac{\epsilon}{2}}{1-\epsilon}D(\mu_2||\mu_2')\right). \tag{2.46}$$

Since $D(\mu_2||\mu_2') > 0$ and $\frac{1-\frac{\epsilon}{2}}{1-\epsilon} > 1$ and because of Proposition 1, we can deduce

$$\lim_{T\to\infty} \mathbb{P}\left(\frac{D(\mu_2||\mu_2')}{(1-\epsilon)\ln T} \times \max_{s\leq f_T} \hat{D}_s(\mu_2||\mu_2')) \leq \frac{1-\frac{\epsilon}{2}}{1-\epsilon}D(\mu_2||\mu_2')\right) = 1. \tag{2.47}$$

As we already know that $\mathbb{P}(C_T) \leq o(1)$ and because of equations (2.45) and (2.46) we can clearly state that

$$\mathbb{P}(N_2(T) < f_n) = o(1). \tag{2.48}$$

Combining this with equation (2.36) we obtain

$$\mathbb{E}[N_2(T)] \geq (1 + o(1))\frac{1-\epsilon}{1+\epsilon}\frac{\ln T}{D(\mu_2||\mu_1)} \tag{2.49}$$

which finishes the proof. □

Theorem 2.2.2 requires the payoffs to be generated by Bernoulli distributions. A generalization of Theorem 2.2.2 can be found in (Lai and Robbins, "Asymptotically Efficient Adaptive Allocation Rules"). We introduce some notation for the general setting. Consider a generic parameters space Θ for a family of rewards distributions $\mathcal{F}$. For any element $\theta \in \Theta$ and $f \in \mathcal{F}$ we write f_θ to mean distribution f parametrized according to θ. Let us assume that we fix a generic element f, then we call its mean when parametrized according to θ as $\mu(\theta)$. We recall the statement of the general Theorem.

Theorem 2.2.3 (Lai and Robbins [1985]). *Assume that $\forall\epsilon > 0$ and $\theta, \lambda \in \Theta$ such that $\mu(\lambda) > \mu(\theta)$, there exists $\delta = \delta(\epsilon, \theta, \lambda) > 0$ so that, if $\mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta$, then $|D(\theta||\lambda) - D(\theta||\lambda')| \leq \epsilon$ and that the parameter space Θ is so that $\forall\lambda \in \Theta$ and $\forall\delta > 0$, there exists $\lambda' \in \theta$ so that $\mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta$.*
Let also π be a policy which,for each fixed parameter vector $\theta = (\theta_1, ..., \theta_K)$, has regret

$$R_\theta(T) = o(n^\alpha) \qquad \text{for every } \alpha > 0. \tag{2.50}$$

Then for every θ such that there exist $k \in \{1, ..., K\}$ so that $\Delta_k \neq 0$ we can state

$$\lim_{T\to\infty} \inf R_\theta(T)/\ln T \geq \sum_{k:\Delta_k>0} \frac{\Delta_k}{D(\theta_k||\theta^*)}. \tag{2.51}$$

Given a policy π and a parameter vector $\theta = (\theta_1, ..., \theta_K)$such that there exist $k \in \{1, ..., K\}$ so that $\Delta_k \neq 0$, if

$$R_\theta(T) \sim \left\{ \sum_{k:\Delta_k>0} \frac{\Delta_k}{D(\theta_k||\theta^*)} \right\} \ln T \qquad \text{with} n \to \infty \tag{2.52}$$

then the policy is said to be *asymptotically efficient*.

2.2.4 Alternative type of policies

In the previous sections we highlighted how the UCB policies are constructed and why they manage to achieve a very low expected regret. However, they are not the only type of policy which can be used to address the stochastic stationary multi-armed bandit problem.
Another famous type of policy is the so called ϵ-greedy. They differ from UCB policy as they don't require the estimation of an upper confidence bound on the estimate of each arm average reward but, instead, they only require an estimate of the average reward itself. Once the average rewards of all arms have been estimated, the one with the highest mean pay-off will be played with probability $1 - \epsilon$ whereas with probability ϵ the decision-maker will play a random arm.

If the quantity ϵ is fixed at the beginning of the game and not changed until the end we are going to give up an average regret of ϵT, which is the expected number of times we would play a suboptimal arm for exploration goals. Therefore, the goal is to find a sequence $(\epsilon_t)_t$ which decreases the more information we have and eventually converges zero.
Specifically, we can define a policy based on this principle called ϵ_t-GREEDY.

Algorithm 3 $\epsilon_t -$ GREEDY

Require: T, K, $c > 0$ and $0 < d < 1$.
Ensure: Play each arm once and define the sequence ϵ_t $(t = 1, 2, ...)$ as

$$\epsilon_t \to \min \left\{ 1, \frac{cK}{d^2 t} \right\}$$

1: **for** $t = 1, 2, ..., T$ **do**
2: Set $\pi_t \leftarrow \underset{k \leq K}{\text{argmax}}\ \hat{x}_k(t)$
3: With probability $1 - \epsilon_t$ play arm π_t whereas with probability ϵ_t play a random arm.

We can notice that, as wanted, the sequence $(\epsilon_t)_t$ decreases linearly with t. The main difference between the UCB-type policies and the ϵ_t-GREEDY is that in the first we tighten the upper confidence the more information we have whereas in the latter we tighten the probability of playing a random arm. More specifically in the UCB strategies if the optimal arm is spotted in the early stages and is played many times at some point the suboptimal arms will have bigger upper confidence bounds which will lead to play them according to exploration-exploitation principle. In the ϵ_t-GREEDY the exploration term is simply given by the sequence $(\epsilon_t)_t$.

Theorem 2.2.4 (Auer et al. [2002a]). *Given a number of arms $K \geq 2$ and the probability distributions $P^{(1)}, ..., P^{(K)}$, if the parameter d of ϵ_t-GREEDY is tuned as*

$$0 < d < \min_{k:\Delta_k \neq 0} \Delta_k$$

then the probability that after t (with $t \geq \frac{cK}{d}$) time steps of the game the algorithm picks a suboptimal arm is bounded by

$$\frac{c}{d^2 t} + 2\left(\frac{c}{d^2}\ln\frac{(t-1)d^2e^{1/2}}{cK}\right)\left(\frac{cK}{(t-1)d^2e^{1/2}}\right)^{\frac{c}{5d^2}} + 4\frac{e}{d^2}\left(\frac{cK}{(t-1)d^2e^{1/2}}\right)^{c/2} \tag{2.53}$$

Proof. Let us introduce the quantity

$$q_0 = \frac{1}{2K}\sum_{s=1}^{t}\epsilon_s \tag{2.54}$$

The probability that the algorithm chooses arm k at time t is

$$\mathbb{P}(\pi_t = k) \leq \left(\frac{\epsilon_t}{K}\right) + \left(1 - \frac{\epsilon_t}{K}\right)\mathbb{P}(\hat{x}_{k,t} \geq \hat{x}_t^*) \tag{2.55}$$

and also

$$\mathbb{P}(\hat{x}_{k,t} \geq \hat{x}_t^*) \leq \mathbb{P}(\hat{x}_{k,t} > \mu_k + \frac{\Delta_k}{2}) + \mathbb{P}(\hat{x}_t^* \leq \mu^* - \frac{\Delta_k}{2}). \tag{2.56}$$

Let us denote by $\tilde{N}_k(t)$ the number of times arm k has been chosen at random in the first t time steps. Then

$$\begin{aligned}
\mathbb{P}(\hat{x}_{k,t} > \mu_k + \frac{\Delta_k}{2}) &= \mathbb{P}(\hat{\mu}_k(N_k(t)) > \mu_k + \frac{\Delta_k}{2}) && (2.57)\\
&= \sum_{s=1}^{t}\mathbb{P}(N_k(t) = s, \hat{\mu}_k(s) > \mu_k + \frac{\Delta_k}{2}) && (2.58)\\
&= \sum_{s=1}^{t}\mathbb{P}\left(N_k(t) = s \mid \hat{\mu}_k(s) > \mu_k + \frac{\Delta_k}{2}\right)\mathbb{P}\left(\hat{\mu}_k(s) > \mu_k + \frac{\Delta_k}{2}\right) && (2.59)\\
&\leq \sum_{s=1}^{t}\mathbb{P}\left(N_k(t) = s \mid \hat{\mu}_k(s) > \mu_k + \frac{\Delta_k}{2}\right)\exp\left(-\frac{\Delta_k^2 s}{2}\right) && (2.60)\\
&\leq \sum_{s=1}^{\lceil q_0\rceil}\mathbb{P}\left(N_k(t) = s \mid \hat{\mu}_k(s) > \mu_k + \frac{\Delta_k}{2}\right) + \frac{2}{\Delta_k^2}\exp\left(-\frac{\Delta_k^2\lceil q_0\rceil}{2}\right) && (2.61)\\
&\leq \sum_{s=1}^{\lceil q_0\rceil}\mathbb{P}\left(\tilde{N}_k(t) \leq s \mid \hat{\mu}_k(s) > \mu_k + \frac{\Delta_k}{2}\right) + \frac{2}{\Delta_k^2}\exp\left(-\frac{\Delta_k^2\lceil q_0\rceil}{2}\right) && (2.62)\\
&\leq q_0 \cdot \mathbb{P}\left(\tilde{N}_k(t) \leq q_0\right) + \frac{2}{\Delta_k^2}\exp\left(-\frac{\Delta_k^2\lceil q_0\rceil}{2}\right) && (2.63)
\end{aligned}$$

where the first inequality follows from the Hoeffding bound, the second comes from dividing the sum in two and bound the first with just the probabilities and the second using $\sum_{s=x+1}^{\infty}\exp(-ks) \leq \frac{1}{k}\exp(-kx)$. The last inequality is caused by the random arm choice being random by definition. Given that

$$\mathbb{E}\left[\tilde{N}_k(t)\right] = \frac{1}{K}\sum_{s=1}^{t}\epsilon_s \tag{2.64}$$

and

$$\mathbb{V}\left[\tilde{N}_k(t)\right] = \sum_{s=1}^{t}\frac{\epsilon_s}{K}\left(1 - \frac{\epsilon_s}{K}\right) \leq \frac{1}{K}\sum_{s=1}^{t}\epsilon_s \tag{2.65}$$

Using Bernstein's Inequality we can state that

$$\mathbb{P}\left(\tilde{N}_k(t) \leq q_0\right) \leq \exp\left(-\frac{q_0}{5}\right). \tag{2.66}$$

Last, we have to lower bound q_0. Given $t \geq t' = \frac{cK}{d^2}$, we can write

$$q_0 = \frac{1}{2K}\sum_{s=1}^{t}\epsilon_s \tag{2.67}$$

$$= \frac{1}{2K}\sum_{s=1}^{t'}\epsilon_s + \frac{1}{2K}\sum_{s=t'+1}^{t}\epsilon_s \tag{2.68}$$

$$\geq \frac{t'}{2K} + \frac{c}{d^2}\ln\frac{n}{n'} \tag{2.69}$$

$$\geq \frac{c}{d^2}\ln\frac{nd^2e^{1/2}}{cK} \tag{2.70}$$

The second term on the right hand side of equation (2.56) can be bounded the same way we did with the first one.
Therefore, we can write

$$\mathbb{P}(\pi_t = k) \geq \frac{\epsilon_t}{K} + 2q_0\exp\left(-\frac{q_0}{5}\right) + \frac{4}{\Delta_k^2}\exp\left(\frac{-\Delta_k^2\lceil q_0\rceil}{2}\right) \tag{2.71}$$

$$\geq \frac{c}{d^2t} + 2\left(\frac{c}{d^2}\ln\frac{(t-1)d^2e^{\frac{1}{2}}}{cK}\right)\left(\frac{cK}{(t-1)d^2e^{1/2}}\right)^{\frac{c}{(5d^2)}} + \frac{4e}{d^2}\left(\frac{cK}{(t-1)d^2e^{1/2}}\right)^{\frac{c}{2}} \tag{2.72}$$

□

2.2.5 KL-UCB

The next strategy we present is called *KL-UCB* belongs again to the *UCB* family and it has been proved to be the only method that always outperforms the classic UCB. This policy has been presented in Garivier and Cappé [2011].
The main difference is in the upper-confidence bound we use to decide which arm we want to play. Specifically, for each arm k at time t, the upper-confidence bound is

$$\max\{q \in \Theta : N_k(t)D(\hat{x}_k(t)||q) \leq \ln(t) + c\ln(\ln(t))\} \tag{2.73}$$

with c parameter. For optimal performances it is recommended to set it equal to 0. This upper-bound returns, for every $k \leq K$, the maximum Bernoulli average q which keeps the K-L Divergence between the average reward at time t for that arm and q times the number of times arm k has been played smaller than $\ln(t) + c\ln(\ln(t))$.
From a computational point of view q can be easily found using Newton iterations, given that for a fixed $p \in [0,1]$ the K-L map $q \to D(p||q)$ is strictly convex and increasing on $[p,1]$.

Algorithm 4 KL-UCB

Require: T, K,c.

Ensure: Play each arm once.

1: **for** $t = 1, 2, ..., T$ **do**

2: Let, $\forall k \leq K$,

$$KLUCB_k \leftarrow \max_{q \in \Theta} N_k(t) D(\hat{x}_k(t)||q) \leq \ln(t) + c \ln(\ln(t))$$

3: Set $\pi_t \leftarrow \underset{k \leq K}{\text{argmax}}\ KL - UCB_k$

4: **end for**

We first present the non-asymptotic bound on the number of times a suboptimal arm can be played.

Theorem 2.2.5 (Garivier and Cappé [2011]). *Let $K \geq 2$ be the number of arms, $\epsilon > 0$ and set the parameter c in Algorithm 4 as $c = 3$. For any positive time t, the number of times Algorithm 4 chooses a suboptimal arm k is bounded by*

$$\mathbb{E}[N_k(t)] \leq \frac{\ln(n)}{D(\mu_k||\mu^*)}(1+\epsilon) + C_1 \ln(\ln(t)) + \frac{C_2(\epsilon)}{n^{\Lambda(\epsilon)}} \tag{2.74}$$

with C_1 positive constant, $C_2(\epsilon)$ and $\Lambda(\epsilon)$ positive functions of ϵ.

Before showing the proof we are going to introduce some results that will help us to go through the proof of Theorem 2.2.5.

Lemma 2.2.1 (Garivier and Cappé [2011]). *For any positive time t*

$$\sum_{s=1}^{t} \mathbb{1}\left\{\pi_s = k, \mu_1 \leq u_1(s)\right\} \leq \sum_{s=1}^{t} \mathbb{1}\left\{sD^+(\hat{x}_k(s)||\mu_1) < \ln(t) + 3\ln(\ln(s))\right\} \tag{2.75}$$

where

$$u_k(t) = \max_{q \geq \hat{x}_k(t)} N_k(t) D(\hat{x}_k(t)||q) \leq \ln(t) + 3\ln(\ln(t)). \tag{2.76}$$

Lemma 2.2.2 (Garivier and Cappé [2011]). *For each $\epsilon > 0$, there is a function $C_2(\epsilon) > 0$ and $\Lambda(\epsilon) > 0$ such that*

$$\sum_{s=W_t+1}^{\infty} \mathbb{P}(D^+(\hat{x}_k(s)||\mu_1) < \frac{D(\mu_k, \mu_1)}{1+\epsilon} \leq \frac{C_2(\epsilon)}{n^{\Lambda(\epsilon)}} \tag{2.77}$$

where

$$W_t = \left\lceil \frac{1+\epsilon}{D^+(\mu_k||\mu_1)} \left(\ln(t) + 3\ln(\ln(t))\right) \right\rceil. \tag{2.78}$$

Theorem 2.2.6. *Let $(X_t)_t \geq 1$ be a sequence of independent identically distributed random variables with support in $[0,1]$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with mean μ. We can define a filtration $\mathcal{F}_t$ such that for any t, $\sigma(X_1, ..., X_t) \subset \mathcal{F}_t$ and for $s > t$, X_s does not depend on $\mathcal{F}_t$. Let $(\epsilon_t)_t$ be a sequence of Bernoulli random variables such that ϵ_t is $\mathcal{F}_{t-1}$-measurable. If we consider a quantity $\delta > 0$, for any $t \in \{1, ..., T\}$, we let*

$$S(t) = \sum_{s=1}^{t} \epsilon_s X_s, \qquad N(t) = \sum_{s=1}^{t} \epsilon_s, \qquad \hat{\mu}(t) = \frac{S(t)}{N(t)}$$

and

$$u(T) = \max\{q > \hat{\mu}_T : N(T)D(\hat{\mu}(T)||q) \leq \delta\}.$$

Then

$$P(u(T) < \mu) \leq e\lceil \delta \ln(T) \rceil \exp(-\delta).$$

Proof of Theorem 2.2.5. Let us assume $\pi^* = 1$ for ease of presentation. As mentioned above, KL-UCB aims to find the arm that maximizes the upper-confidence bound

$$u_k(t) = \max_{q \geq \hat{x}_k(t)} N_k(t)D(\hat{x}_k(t)||q) \leq \ln(t) + 3\ln(\ln(t)). \tag{2.79}$$

We define $D^+(x||y) = D(x||y)\mathbb{1}\{x < y\}$ for $x, y \in [0, 1]$. We bound the expected number of times a suboptimal arm is played as

$$\mathbb{E}[N_k(t)] = \mathbb{E}\left[\sum_{s=1}^{t} \mathbb{1}_{\{}\pi_s = k\}\right] \leq \mathbb{E}\left[\sum_{s=1}^{t} \mathbb{1}\{\mu_1 > u_1(s)\}\right] + \mathbb{E}\left[\sum_{s=1}^{t} \mathbb{1}\{\pi_s = k, \mu_1 \leq u_1(s)\}\right] \tag{2.80}$$

$$\leq \sum_{s=1}^{t} \mathbb{P}(\mu_1 > u_1(s)) + \mathbb{E}\left[\sum_{s=1}^{t} \mathbb{1}\{sD^+(\hat{x}_k(s)||\mu_1) < \ln(t) + 3\ln(\ln(t))\}\right] \tag{2.81}$$

with the last inequality following from Lemma 1. The first term of line (2.77) can be bounded using Theorem 6. Thus,

$$\mathbb{P}(\mu_1 > u_1(s)) \leq e\lceil \ln(s)(\ln(s)) + 3\ln(\ln(s)))\rceil \exp(-\ln(s) - 3\ln(\ln(s))) \tag{2.82}$$

$$= e\frac{\lceil \ln(s)^2 + 3\ln(s)\ln(\ln(s))\rceil}{s\ln(s)^3}. \tag{2.83}$$

Therefore, we can rewrite

$$\sum_{s=1}^{t} \mathbb{P}(\mu_1 > u_1(s)) \leq \sum_{s=1}^{t} \frac{e\lceil \ln(s)^2 + 3\ln(s)\ln(\ln(s))\rceil}{s\ln(s)^3} \leq C_1 \ln(\ln(t)) \tag{2.84}$$

for some positive constant C_1. In the original paper it is said that $C_1 \leq 7$ would be an adequate value. To bound the second term of line (2.77) we first define a quantity W_t such that

$$W_t = \left\lceil \frac{1+\epsilon}{D^+(\mu_k||\mu_1)} \left(\ln(t) + 3\ln(\ln(t))\right) \right\rceil. \tag{2.85}$$

This leads to

$$\sum_{s=1}^{t} \mathbb{P}(sD^+(\hat{x}_k(s)||\mu_1) < \ln(t) + 3\ln(\ln(t))) \tag{2.86}$$

$$\leq W_t + \sum_{s=W_t+1}^{\infty} \mathbb{P}(sD^+(\hat{x}_k(s)||\mu_1) < \ln(t) + 3\ln(\ln(t))) \tag{2.87}$$

$$\leq W_t + \sum_{s=W_t+1}^{\infty} \mathbb{P}(W_tD^+(\hat{x}_k(s)||\mu_1) < \ln(t) + 3\ln(\ln(t))) \tag{2.88}$$

$$= W_t + \sum_{s=W_t+1}^{\infty} \mathbb{P}(D^+(\hat{x}_k(s)||\mu_1) < \frac{D^(\mu_k, \mu_1)}{1+\epsilon}) \tag{2.89}$$

$$\leq \frac{1+\epsilon}{D^+(\mu_k||\mu_1)} \left(\ln(t) + 3\ln(\ln(t))\right) + \frac{C_2(\epsilon)}{n^{\Lambda(\epsilon)}} \tag{2.90}$$

according to Lemma 2.2.2. This finishes the proof. □

2.3 Strategies for Adversarial Bandits

We now go back to the adversarial setting defined previously in the work. Recall that, unlike the stochastic one, in this setting the rewards are not assumed to be generated by any probability distribution but we assume that they are set from the environment at each time step t. In this setting the regret incurred in T rounds is defined as

$$R(T) = \max_{k \leq K} \sum_{t=1}^{T} x_k(t) - \sum_{t=1}^{T} x_{\pi_t}(t)$$

where $x_k(t)$ is the reward obtained by playing arm k at time t and π_t is the arm chosen by the used policy at time t. Recall that the rewards in this setting are not denoted with an upper case letter because they are not random.
A symmetric definition of the regret can be given using losses rather than gains. Specifically, define $l_k(t) = 1 - x_k(t)$, then we can rewrite the regret as

$$R(T) = \sum_{t=1}^{T} l_{\pi_t}(t) - \min_{k \leq K} \sum_{t=1}^{T} l_k(t).$$

In this section we will mainly use this version of the regret.
Like the stochastic setting, we want to find a policy that can obtain a sublinear regret. However, if the rewards are non-stochastic, this goal is harder as we cannot learn like we did in the stochastic case. As the environment might be either *nonoblivious* either *oblivious*, he might put lower rewards in the arms we are more likely to play. Thus, the goal of the decision maker has to be to surprise the environment. To achieve this we have to add randomness to the arm chosen by the policy so that the adversary cannot predict the action of the forecaster as easily.
We are going to investigate the pseudo-regret which is defined as

$$\overline{R}_T = \mathbb{E} \sum_{t=1}^{T} l_{\pi_t}(t) - \min_{k \leq K} \mathbb{E} \sum_{t=1}^{T} l_k(t). \tag{2.91}$$

The policy we are going to study is called Exp3 and uses the randomization factor mentioned above. This strategy has been described for the first time in Auer et al. [2002b].
One of the main features of this algorithm is its ability to construct unbiased estimators for the loss of each arm. Specifically, suppose that the decision-maker chooses which arm to play next π_t according to a probability distribution $p_t = (p_t^1, ..., p_t^k)$, then we can define an unbiased estimator of $l_k(t)$ as

$$\tilde{l}_k(t) = \frac{l_k(t)}{p_t^k} \mathbb{1}_{\pi_t = k}. \tag{2.92}$$

The second intuition used in this algorithm is to set the above mentioned probabilities $p_t = (p_t^1, ..., p_t^k)$ using an exponential reweighting of the cumulative estimated loss of each arm. The algorithm follows below.

The regret analysis of this algorithm below shows that Exp3 successfully achieves a sublinear regret.

Theorem 2.3.1 ([Auer et al., 2002b]). *Let $\eta_t = \eta = \frac{2 \ln K}{TK}$ for the input parameter of Exp3. Then*

$$\overline{R}_T \leq \sqrt{2TKlnK}.$$

If $\eta_t = \frac{2 \ln K}{tK}$ then

$$\overline{R}_T \leq 2\sqrt{TKlnK}.$$

Algorithm 5 Exp3: Exponential weights for Exploration and Exploitation Auer et al. [2002b]

Require: $(\eta_t)_{t\in\mathcal{N}}$ nonincreasing sequence of real number.

1: **for** $t = 1, 2, ..., T$ **do**
2: Draw an arm π_t from p_t.
3: Compute the estimated loss $\tilde{l}_k(t) = \frac{l_k(t)}{p_t^k}\mathbb{1}_{\pi_t=k}$ and update the cumulative one $\tilde{L}_k(t) = \tilde{L}_k(t-1) + \tilde{l}_k(t)$.
4: Compute the probability distribution for the next round $p_{t+1} = (p_{t+1}^1, ..., p_{t+1}^k)$ with

$$p_{t+1}^k = \frac{\exp(-\eta_t \tilde{L}_k(t))}{\sum_{k'=1}^K \exp(-\eta_t \tilde{L}_{k'}(t))}$$

Proof of Theorem 7. We want to prove that, given any nonincreasing real sequence $(\eta_t)_{t\in\mathcal{N}}$, the pseudo-regret of Exp3 is bounded as

$$\overline{R}_T \leq \frac{K}{2}\sum_{t=1}^T \eta_t + \frac{\ln K}{\eta_T}. \tag{2.93}$$

Following Equation (2.92) we can write

$$\mathbb{E}_{\pi_t\sim p_t}[\tilde{l}_k(t)] = \sum_{j=1}^K p_t^j \frac{l_k(t)}{p_t^k}\mathbb{1}\{\pi_t = k\}. \tag{2.94}$$

The equalities below can be easily derived from (2.94):

$$\mathbb{E}_{k\sim p_t}[\tilde{l}_k(t)] = l_{\pi_t}(t), \quad \mathbb{E}_{\pi_t\sim p_t}[\tilde{l}_k] = l_{k_t}(t), \quad \mathbb{E}_{\pi_t\sim p_t}[\tilde{l^2}_k(t)] = \frac{l_{\pi_t}^2(t)}{p_t^{\pi_t}}, \quad \mathbb{E}_{\pi_t\sim p_t}\left[\frac{1}{p_t^{\pi_t}}\right] = K. \tag{2.95}$$

Therefore, they imply

$$\sum_{t=1}^T l_{\pi_t}(t) - \sum_{t=1}^T l_{k'}(t) = \sum_{t=1}^T \mathbb{E}_{k\sim p_t}[\tilde{l}_{k'}(t)] - \sum_{t=1}^T \mathbb{E}_{\pi_t\sim p_t}[\tilde{l}_{k'}(t)]. \tag{2.96}$$

We can rewrite $\mathbb{E}_{k\sim p_t}[\tilde{l}_k(t)]$ as

$$\mathbb{E}_{k\sim p_t}[\tilde{l}_k(t)] = \frac{1}{\eta_t}\ln \mathbb{E}_{k\sim p_t}\exp\left(-\eta_t(\tilde{l}_k(t) - \mathbb{E}_{k'\sim p_t}[\tilde{l}_{k'}(t)])\right) - \frac{1}{\eta_t}\ln \mathbb{E}_{k\sim p_t}\exp\left(-\eta_t\tilde{l}_k(t)\right) \tag{2.97}$$

where $\ln \mathbb{E}_{k\sim p_t}\exp\left(-\eta_t\tilde{l}_k(t)\right)$ is the logarithm of the moment generating function of $\tilde{l}_k(t)$. The next part of the proof is going to deal with bounding the terms of Equation (2.97).
We start with the first one:

$$\begin{aligned}
\ln \mathbb{E}_{k\sim p_t}\exp\left(-\eta_t(\tilde{l}_k(t) - \mathbb{E}_{k'\sim p_t}[\tilde{l}_{k'}(t)])\right) &= \ln \mathbb{E}_{k\sim p_t}\exp\left(-\eta_t\tilde{l}_k(t)\right) + \eta_t\mathbb{E}_{k'\sim p_t}[\tilde{l}_{k'}(t)] && (2.98)\\
&\leq \mathbb{E}_{k\sim p_t}\left(\exp\left(-\eta_t\tilde{l}_k(t)\right) - 1 + \eta_t\tilde{l}_{k'}(t)\right) && (2.99)\\
&\leq \mathbb{E}_{k\sim p_t}\left[\frac{\eta_t^2\tilde{l}_k^2(t)}{2}\right] && (2.100)\\
&\leq \frac{\eta_t^2}{2p_t^{\pi_t}} && (2.101)
\end{aligned}$$

where the first inequality follows from the well know inequality $\ln x \leq x - 1$, the second from $\exp(-x) - 1 + x \leq \frac{x^2}{2}$ and the last inequality comes from the third term of Equation (2.95).

Concerning the second term of the right-hand side of Equation (2.97), we define $\tilde{L}_k(0) = 0,\ \phi_0(\eta) = 0$ and $\phi_t(\eta) = \frac{1}{\eta} \ln \frac{1}{K} \sum_{k=1}^{K} \exp(-\eta_t \tilde{L}_k(t))$. We use the definition of p_t to write

$$-\frac{1}{\eta_t} \ln \mathbb{E}_{k\sim p_t} \exp\left(-\eta_t \tilde{l}_k(t)\right) = -\frac{1}{\eta_t} \ln \frac{\sum_{k=1}^{K} \exp(-\eta_t \tilde{L}_k(t))}{\sum_{k=1}^{K} \exp(-\eta_t \tilde{L}_k(t-1))} = \phi_{t-1}(\eta_t) - \phi_t(\eta_t). \tag{2.102}$$

Using results from Equation (2.97), (2.101) and (2-102) in Equation (2.96) we get

$$\sum_{t=1}^{T} x_k(t) - \sum_{t=1}^{T} x_{\pi_t}(t) \leq \sum_{t=1}^{T} \frac{\eta_t}{2p_t^{\pi_t}} + \sum_{t=1}^{T} (\phi_{t-1}(\eta_t) - \phi_t(\eta_t)) - \sum_{t=1}^{T} \mathbb{E}_{k\sim p_t}[\tilde{l}_k(t)] \tag{2.103}$$

Because of the last term in Equation (2.96) and because of the properties of conditional expectation we can bound the first term of the right-hand side as

$$\mathbb{E} \sum_{t=1}^{T} \frac{\eta_t}{2p_t^{\pi_t}} = \mathbb{E} \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p_t} \frac{\eta_t}{2p_t^{\pi_t}} = \frac{K}{2} \sum_{t=1}^{T} \eta_t. \tag{2.104}$$

We deal with the second term first using an Abel transformation as, given that $\phi_0(\eta_1) = 0$,

$$\sum_{t=1}^{T} (\phi_{t-1}(\eta_t) - \phi_t(\eta_t)) = \sum_{t=1}^{T-1} (\phi_t(\eta_{t+1}) - \phi_t(\eta_t)) - \phi_T(\eta_T). \tag{2.105}$$

Thus,

$$-\phi_T(\eta_T) = \frac{\ln K}{\eta_T} - \frac{1}{\eta_T} \ln \left(\sum_{k'=1}^{K} \exp(\eta_T \tilde{L}_{k'}(T)) \right) \tag{2.106}$$

$$\leq \frac{\ln K}{\eta_T} - \frac{1}{\eta_T} \ln \left(\exp(\eta_T \tilde{L}_k(T)) \right) \tag{2.107}$$

$$= \frac{\ln K}{\eta_T} + \sum_{t=1}^{T} \tilde{l}_k(t) \tag{2.108}$$

and therefore

$$\mathbb{E}\left[\sum_{t=1}^{T} x_k(t) - \sum_{t=1}^{T} x_{\pi_t}(t) \right] \leq \frac{K}{2} \sum_{t=1}^{T} \eta_t + \frac{\ln K}{\eta_T} + \mathbb{E}\left[\sum_{t=1}^{T-1} (\phi_t(\eta_{t+1}) - \phi_t(\eta_t)) \right]. \tag{2.109}$$

We just need to show that the function $\phi_t(\eta)$ is increasing and thus $\phi_t'(\eta) \geq 0$. We know that $(\eta_t)_t$ is a non-increasing sequence, so we have $\eta_{t+1} \leq \eta_t$. Therefore, we would to obtain $\phi_t(\eta_{t+1} - \eta_t)$. Define,

$$p_t^k(\eta) = \frac{\exp(-\eta \tilde{L}_k(t))}{\sum_{k'=1}^{K} \exp(-\eta \tilde{L}_{k'}(t))}. \tag{2.110}$$

Then,

$$\phi_t'(\eta) = -\frac{1}{\eta^2} \ln \left(\frac{1}{K} \sum_{k=1}^{K} \exp(-\eta \tilde{L}_k(t)) \right) - \frac{1}{\eta} \frac{\sum_{k=1}^{K} \tilde{L}_k(t) \exp(-\eta \tilde{L}_k(t))}{\sum_{k=1}^{K} \exp(-\eta \tilde{L}_k(t))} \tag{2.111}$$

$$= \frac{1}{\eta^2} \frac{1}{\eta^2 \sum_{k=1}^{K} \exp(-\eta \tilde{L}_k(t))} \sum_{k=1}^{K} \exp(-\eta \tilde{L}_k(t)) \tag{2.112}$$

$$\times \left(-\eta \tilde{L}_k(t) - \ln \left(\frac{1}{K} \sum_{k=1}^{K} \exp(-\eta \tilde{L}_k(t)) \right) \right). \tag{2.113}$$

To conclude, doing simplifications returns

$$\phi_t^{'}(\eta) = \frac{1}{\eta^2}\sum_{k=1}^{K} p_t^k(\eta)\ln(Kp_t^k(\eta)) = \frac{1}{\eta^2}D(p_t^k(\eta)||p^1). \tag{2.114}$$

As the Kullback-Leibler Divergence between two probability distribution is always greater or equal than zero it concludes the proof. □

Therefore, the strategy Exp3 for Adversarial Multi-Armed Problems achieves a pseudo-regret of order $O(\sqrt{T})$. However, even if the pseudo-regret is sublinear, it is not as optimal as the logarithmic one found for the UCB-like policies in the stochastic stationary setting. This is somehow expected as the assumption made in the adversarial setting are way milder and thus the problem is intrinsically harder.
We now move back to the stochastic setting with different assumptions.

Chapter 3

Stochastics Bandits: Non-Stationary Case

In this subsection we deal again with the Multi-Armed Bandit Problem, however the assumptions we make are different. Specifically, we no longer assume that the reward distribution on each arm has to remain the same for the whole duration of the game. This setting is usually referred to as the *non-stationary case*.
The element of *non-stationarity* that will be brought reflects what we would need if we tried to address some real-world problems using bandits. In facts, assuming that the expected reward the forecaster would get playing one arm is going to remain the same for the whole duration of the game is something that we would not be necessairily abe to do everytime. In most real worlds scenarios, in fact, the expected reward the decision-maker obtains when making a choice, will most likely vary through time: choosing to invest on goods useful for the winter will probably give me more profit during the winter than during the summer or, from a more financial point of view, if I have to put money on a financial security it would be useful to choose the one that has the highest expected reward in this period, knowing that its value might have been different previously.

The notion of *non-stationarity* in this problem can give rise to a range of different settings, from piece-wise stationarity of the underlying processes to the assumption that the reward distributions have some form of drift. The setting considered by Garivier and Moulines [2011] corresponds to the case where each arm is modelled by a sequence of independent random variables with bounded support such that the joint distribution of the consecutive random variables undergoes abrupt changes at unknown time-instants called *break-points* or *change-points*. By contrast, a continuously changing environment is modelled by Slivkins and Upfal [2008] where each arm independently follows a Brownian motion.

The regret in the piece-wise stationary setting is defined with respect to the optimal strategy that tracks the best arm at every time-step. In this case, Garivier and Moulines [2011] derive an $\mathcal{O}(\sqrt{T})$ lower-bound on the regret after T rounds of play and provide an $\mathcal{O}(T^{\beta+1}\log T)$ upper-bound on the regret of the so-called D(iscounted)-UCB algorithm proposed by Kocsis L. [2006], under the assumption that the number of change-points is $\mathcal{O}(T^{\beta})$ for some $\beta \in [0, 1)$, and provided that the discount factor is appropriately tuned. An alternative UCB-type approach proposed by Garivier and Moulines [2011] is based on a sliding window, which is shown to achieve a slightly better regret bound, for an appropriate choice of window size as a function of the number change-points.

Cao et al. [2019] considers the case where there is a known number M of piece-wise stationary segments. The length of each stationary segment is at least L, which as follows from Assumption 3.4.1 (taken from the above mentioned paper) , is known to the algorithm. Moreover, it is assumed that the change amplitude is over a certain threshold for at least one arm at each change-point. This

assumption is made as this algorithm uses a change-point detector to spot changes in the rewards distribution of arms.
Before moving forward with the description of the above mentioned algorithms we introduce and redefine some notations and concepts.

3.1 Formulation of the Non-Stationary case

Let K be the number of arms in the problem like in the previous sections. We define **change-points** the times on the timeline $[1, T]$ where the rewards distributions change. More specifically, if we assume to have $m << T$ change-points equals for all the arms then we can denote them as $\nu_1, ..., \nu_m$ with $\nu_1 < \nu_2 < ... < \nu_m$. The case we just described relies on the assumption that we can't have different arms having changes at different times, we therefore call this type of changes **global**.
If we assume that each arm can change rewards distribution m times but separately from the others then we have to introduce the notation $\nu_1^{(k)}, ..., \nu_m^{(k)}$ where $\nu_i^{(k)}$ indicates the **i**-th change on the **k**-th arm where k $\in \{1, ..., K\}$. In this case we call the change-points **local**.
In this work we will mostly deal with global change-points.

The policies π_t are defined the same way as before as strategies that aim to spot the optimal arm based on past rewards from all arms.
Let $\mu_t^* = \mathbb{E}[X^*(t)]$ be the highest mean among all the means of the K arms at time T. Therefore we can define the expected regret incurred using policy π at time T as

$$\mathbb{E}[R_\pi(T)] = \sum_{t=1}^{T} \mu_t^* - \sum_{t=1}^{T} \mathbb{E}\left[X_{\pi_t}(t)\right] \tag{3.1}$$

If we strengthen the assumptions letting the rewards being in $[0, 1]$ then the expected regret becomes the expected amount of times the forecaster will play a suboptimal arm following policy π.
We now show two similar algorithms designed for the piecewise-stationary setting with abruptly changing environments.

3.2 Discounted UCB and Sliding-Window UCB

Discounted UCB (D-UCB) and Sliding-Window UCB (SW-UCB) are two algorithms presented in [Garivier and Moulines, 2011] with the goal of achieving a sub-linear regret on the piecewise-stationary with abruptly changing environments multi-armed bandit. They are based on the similar concept of penalising past observation. However, while one uses a milder approach which consists in penalising more older observations (D-UCB) the second uses a more drastic approach based on not penalising the last τ observations and forgetting all the other observations.
The urge to change the policies designed for the stochastic stationary setting followed Hartland [2006] who showed that empirically the strategies previously created were not appropriate for this new setting.
The two above mentioned algorithms are variations of the classic UCB approach. D-UCB is based on the idea of introducing a discount factor $\gamma \in (0, 1)$ to penalise the older observations in an exponential way. This factor has been introduced to reduce the bias in the estimates of the expected pay-off of each arm, where the bias is brought from the observations coming from old rewards distributions.
Specifically, for the D-UCB we can redefine the expected pay-off of arm k at time t with a discount

γ as

$$\hat{x}_k(t,\gamma) = \frac{1}{N_k(t,\gamma)} \sum_{s=1}^{t} \gamma^{t-s} x_k(s) \mathbb{1}\{\pi_s = k\} \tag{3.2}$$

where

$$N_k(t,\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}\{\pi_s = k\}. \tag{3.3}$$

As the D-UCB algorithm belongs to the UCB family, the criterion used to choose the optimal arm is still the maximization of the quantity $\hat{x}_k(t,\gamma) + c_k(t,\gamma)$ with $c_k(t,\gamma)$ discounted exploration bound defined as

$$c_k(t,\gamma) = 2\sqrt{\frac{\xi \ln(n_t(\gamma))}{N_k(t,\gamma)}} \tag{3.4}$$

with $n_t(\gamma) = \sum_{k=1}^{K} N_k(t,\gamma)$, the rewards in the interval $[0,1]$ and an appropriate choice of ξ.
We can notice that in the D-UCB algorithm, if the parameter γ is set to be either 0 or 1, yields

Algorithm 6 D-UCB

Require: T, K, γ, ξ.
Ensure: Play each arm once.

1: **for** $t = 1, 2, ..., T$ **do**
2: Let, $\forall k \leq K$,
$$DUCB_k \leftarrow \hat{x}_k(t,\gamma) + c_k(t,\gamma)$$
3: Set $\pi_t \leftarrow \underset{k \leq K}{argmax} DUCB_k$
4: **end for**

extreme results. Specifically if $\gamma = 0$ then every arm would have expected pay-off equal to 0 whereas if $\gamma = 1$ we would go back to the classic UCB strategy.
Before moving to a more detailed analysis of D-UCB we present the SW-UCB algorithm. The SW-UCB is based on computing the averages on a fixed-size horizon τ. This can be seen as a D-UCB where the discount is set to 1 for the τ latest observations while is set to 0 for the older rewards. The expected pay-off for arm k at time t with the window of size τ is

$$\hat{x}_k(t,\tau) = \frac{1}{N_k(t,\tau)} \sum_{s=t-\tau+1}^{t} x_k(s) \mathbb{1}\{\pi_s = k\} \tag{3.5}$$

and

$$N_k(t,\tau) = \sum_{s=t-\tau+1}^{t} \mathbb{1}\{\pi_s = k\}. \tag{3.6}$$

For this algorithm the window exploration bound is

$$c_k(t,\tau) = \sqrt{\frac{\xi \ln(\min(t,\tau))}{N_k(t,\tau)}} \tag{3.7}$$

if the rewards are in the interval $[0,1]$ and ξ tuned appropriately. Let M_T be the number of changepoints that happen by time T and let $\overline{N}_k(T)$ be the number of times arm k has been played

Algorithm 7 SW-UCB

Require: T, K, τ, ξ.
Ensure: Play each arm once.

1: **for** $t = 1, 2, ..., T$ **do**
2: Let, $\forall k \leq K$,
$$SWUCB_k \leftarrow \hat{x}_k(t, \tau) + c_k(t, \tau)$$
3: Set $\pi_t \leftarrow \underset{k \leq K}{argmax}\ SWUCB_k$
4: **end for**

while being suboptimal. Recall that in this setting arm k might be suboptimal in some segments of the timeline and optimal in others.
For the D-UCB algorithm, given the definition of Regret we gave with Equation (3.1), we can write

$$\mathbb{E}[R_\pi(T)] = \sum_{t=1}^{T} \mu_t^* - \sum_{t=1}^{T} \mathbb{E}\left[X_{\pi_t}(t)\right] \leq \sum_{k=1}^{K} \mathbb{E}_\gamma\left[\overline{N}_k(T)\right] \tag{3.8}$$

for rewards in $[0, 1]$, where $\mathbb{E}_\gamma$ is the expectation evaluated w.r.t. the D-UCB strategy with discount factor γ.
For the SW-UCB we can use a similar bound where the expectation $\mathbb{E}_\gamma$ is calculated with respect to the SW-UCB policy with window size τ and is therefore denoted as $\mathbb{E}_\tau$. Therefore, similarly to the previous cases, our goal is to upperbound the number of times a suboptimal arm has been played.
As the two algorithms are very similar we are going to show the bound for both algorithms but the regret analysis will be shown for the D-UCB algorithm only.

Theorem 3.2.1 (Garivier and Moulines [2011]). *Let the input parameters of the D-UCB policy be $\xi \in (1/2, 1)$ and $\gamma \in (1/2, 1)$. Then,*

$$\mathbb{E}_\gamma[\overline{N}_k(T)] \leq C_1 T(1-\gamma)\ln\frac{1}{1-\gamma} + C_2\frac{M_T}{1-\gamma}\ln\frac{1}{1-\gamma} \tag{3.9}$$

with

$$C_1 = \frac{32\sqrt{2}B^2\xi}{\gamma^{1/(1-\gamma)}(\Delta_{\mu_T}(k))^2} + \frac{4}{(1-\ln\frac{1}{e})\ln\left(1+4\sqrt{1-1/2\xi}\right)} \tag{3.10}$$

and

$$C_2 = \frac{\gamma-1}{\ln(1-\gamma)\ln(\gamma)} \times \ln((1-\gamma)\xi\ln(n_T(\gamma))). \tag{3.11}$$

We can say that for γ going to 1, and thus for a policy close to the UCB classic one and $C_2 \to 1$, we have

$$C_1 \to \frac{16eB^2\xi}{(\Delta_{\mu_T}(k))^2} + \frac{2}{(1-e^{-1})\ln\left(1+4\sqrt{1-1/2\xi}\right)} \tag{3.12}$$

As mentioned above, we report the bound for the SW-UCB as well.

Theorem 3.2.2 (Garivier and Moulines [2011]). *Let the input parameters of the SW-UCB policy be $\xi \geq \frac{1}{2}$ and $\tau \in \mathbb{N}$. Then,*

$$\mathbb{E}_\gamma[\overline{N}_k(T)] \leq C(\tau)\frac{T\ln(\tau)}{\tau} + \tau M_T + \ln^2(\tau) \tag{3.13}$$

with

$$C(\tau) = \frac{4B^2\xi}{(\Delta_{\mu_T}(k))^2}\frac{\lceil T/\tau \rceil}{T/\tau} + \frac{2}{\ln(\tau)}\left\lceil \frac{\ln(\tau)}{\ln(1+4\sqrt{1-1/2\xi})} \right\rceil. \tag{3.14}$$

If we let $\tau \to \infty$ *and* $\frac{T}{\tau} \to \infty$ *we get*

$$C(\tau) \to \frac{4B^2\xi}{(\Delta_{\mu_T}(k))^2} + \frac{2}{\ln(1+4\sqrt{1-1/2\xi})} \tag{3.15}$$

Before showing the proof of Theorem 8 we introduce a Theorem and Lemma that we will use in the proof itself.

Theorem 3.2.3 (Garivier and Moulines [2011]). *Let* $(X_t)_t$ *be a sequence of independent random variables so that* $\mathbb{E}[X_t] = \mu_t$ *and* $(\epsilon)_t$ *be a sequence of independent Bernoulli random variables .For all integers* t *and* $\delta, \eta > 0$*, if we define* $S_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} X_s \epsilon_s$*,* $G_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \mu_s \epsilon_s$*,* $N_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s$ *and* $n_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s$*. Then, we can state*

$$\mathbb{P}\left(\frac{S_t(\gamma) - G_t(\gamma)}{\sqrt{N_t(\gamma^2)}} \geq \delta \right) \leq \left\lceil \frac{\ln(n_t(\gamma))}{\ln(1+\eta)} \right\rceil \exp\left(-\frac{2\delta^2}{B^2}\left(1 - \frac{\eta^2}{16}\right)\right). \tag{3.16}$$

This teorem bounds the probability of the sequence of rewards X_t being $\delta\sqrt{N_t(\gamma^2)}$ far apart from the true mean. It can be seen as a weaker Hoeffding-Chernoff bound for a scenario with discount.

Lemma 3.2.1 (Garivier and Moulines [2011]). *Let* $k \leq K$ *and* τ *integer. Let* $N_k(t-\tau : t) = \sum_{s=t-\tau+1}^{t} \mathbb{1}\{\pi_t = k\}$*. Then, for* $m > 0$*,*

$$\sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k, N_k(t-\tau : t) < m\} \leq \left\lceil \frac{T}{\tau} \right\rceil m \tag{3.17}$$

which implies that for any $\tau \geq 1$ *and* $A > 0$ *we have* $\sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k, N_k(t-\tau : t, \gamma) < A\} \leq \lceil \frac{T}{\tau} \rceil A\gamma^{-\tau}$*.*

This lemma provides a bound on the number of times a suboptimal arm can be played by a policy with rewards discount γ and with window size τ. In case we set $\tau = T$ we obtain the special case of the lemma for the D-UCB policy only whereas if we let $\gamma = 1$ the lemma becomes designed for the SW-UCB strategy.
Using these two results, we can move to the proof of Theorem 3.2.1.

Proof of Theorem 3.2.1. The first thing we have to consider when moving into this proof is that in this scenario $\hat{x}_k(t,\gamma)$ is going to provide a biased estimate of the mean of arm k at time t $\mu_k(t)$ and therefore we will have to address how big the bias is. Another component that will differ from the rest of the UCB proofs we have seen up to here is that instead of the Chernoff-Hoeffding we will use the concentration inequality defined in Theorem 3.2.3.

We will follow the structure given to the proof by [Garivier and Moulines, 2011] where the authors divided the proof in some steps.

Step 1: In the first step the main focus is going to be on dividing the number of times arm k has been played while being suboptimal before time t, $\overline{N}_k(t)$, into quantities than can be studied and bounded in an easier way. Specifically,

$$\overline{N}_k(T) = 1 + \sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k \neq k^*, N_k(t,\gamma) < Q(\gamma)\} + \sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k \neq k^*, N_k(t,\gamma) \geq Q(\gamma)\} \tag{3.18}$$

with $Q(\gamma) = 16\xi \ln(n_T(\gamma))/(\Delta_{\mu_T}(k))^2$. To upper-bound the first term in the expression above we can use Lemma 3.2.1 which allows us to write

$$\sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k \neq k^*, N_k(t,\gamma) < Q(\gamma)\} \leq \lceil T(1-\gamma)\rceil Q(\gamma)\gamma^{-\frac{1}{1-\gamma}}. \tag{3.19}$$

When a breakpoint occurs, for a number of rounds depending on how fast we forget of the past (i.e. how small γ is), we are going to have very biased expected rewards estimates. This number of time steps with poor estimates is defined as $R(\gamma) = \ln((1-\gamma)\xi \ln(n_K(\gamma)))/\ln(\gamma)$ rounds. For any positive time horizon T we define the set $J(\gamma)$ of the time steps $t \in \{K+1, ..., T\}$ so that for all $s \in (t - R(\gamma), t)$ and for every arm $k \leq K$ we have $\mu_s(k) = \mu_t(k)$. This basically represents the set of time indices such that they are not to close to a change in distribution. Thus, we can rewrite

$$\sum_{t=K+1}^{T} \mathbb{1}\{\pi_t = k \neq k^*, N_k(t,\gamma) \geq Q(\gamma) \leq M_T R(\gamma) + \sum_{t \in J(\gamma)} \mathbb{1}\{\pi_t = k \neq \pi_t^*, N_k(t,\gamma) \geq Q(\gamma)\}. \tag{3.20}$$

Therefore, we have now a bound on the most problematic phases of the game as we have previously bounded the amount of times the decision maker chooses the suboptimal arm k when the information available is not much and after a change. In the second step we will deal with the last quantity to bound.

Step 2: The approach we are going to use goes back to the classic one used in UCB analysis where in order to have a suboptimal arm played we either have to have the optimal arm heavily underestimated, the suboptimal one heavily overestimated or the means of the optimal and suboptimal ones very close to each other.
Let $t \in J(\gamma)$. Formalizing what we mentioned above, we need one of the following conditions

$$\begin{cases} \hat{x}_k(t,\gamma) + c_k(t,\gamma) < \mu_k(t) + 2c_k(t,\gamma) \\ \hat{x}_k(t,\gamma) + c_k(t,\gamma) < \mu_t^* \\ \mu_t^* < \hat{x}^*(t,\gamma) + c^*(t,\gamma) \end{cases} \tag{3.21}$$

has to be false in order to have the suboptimal arm k played. If all of them were true then we would have $\hat{x}_k(t,\gamma) + c_k(t,\gamma) < \hat{x}^*(t,\gamma) + c^*(t,\gamma)$.
Therefore,

$$\{\pi_t = k \neq \pi_t^*, N_k(t,\gamma) \geq Q(\gamma)\} \subseteq \begin{cases} \{\mu_t^* - \mu_k(t) < 2c_k(t,\gamma), N_k(t,\gamma) \geq Q(\gamma)\} \\ \cup\{\hat{x}^*(t,\gamma) \leq \mu^*(t) - c^*(t,\gamma)\} \\ \cup\{\hat{x}_k(t,\gamma) \geq \mu_k(t) - c_k(t,\gamma)\}. \end{cases} \tag{3.22}$$

However, by the definition we gave of $Q(\gamma)$, we know that $c_k(t,\gamma) \leq 2\sqrt{\xi \ln(n_t(\gamma))/Q(\gamma)} \leq \Delta_{\mu_T}(k)/2$ which implies that the first inequality in the system above cannot occur.
Our goal is now to upper-bound the probabilities of the last two events of system (139). The main idea is to show that the behaviour of the estimates after at least $R(\gamma)$ rounds after a change tends to be good and therefore the probability of a suboptimal arm being played is very low.
Let us define the event $\mathcal{E}_k(t,\gamma) = \{\hat{x}_k(t,\gamma) \geq \mu_k(t) + c_k(t,\gamma)\}$.
We now have to bound the probability of $\mathcal{E}_k(t,\gamma)$ by considering both the fluctuations of the expected rewards $\hat{x}_k(t,\gamma)$ around $\frac{V_k(t,\gamma)}{N_k(t,\gamma)}$, where $V_k(t,\gamma) = \sum_{s=1}^{t} \gamma^{t-s}\mu_k(t)\mathbb{1}\{\pi_s = k\}$, and the bias is $\frac{V_k(t,\gamma)}{N_k(t,\gamma)} - \mu_k(t)$.

Step 3: We first analyse the bias. We can notice $\frac{V_k(t,\gamma)}{N_k(t,\gamma)}$ is a convex combinations of the possible means $\mu_k(s)$ with $s \in \{1, ..., t\}$. Thus, as the rewards are in $[0,1]$ then $|V_k(t,\gamma)/N_k(t,\gamma) -$

$\mu_k(t)|\leq 1$. For every 'good' t (e.g. $t \in J(\gamma)$), we have

$$|V_k(t,\gamma) - \mu_k(t)N_k(t,\gamma)| = \left| \sum_{s=1}^{t-R(\gamma)} \gamma^{t-s}(\mu_k(s) - \mu_k(t))\mathbb{1}\{\pi = t\} \right| \tag{3.23}$$

$$\leq \sum_{s=1}^{t-R(\gamma)} \gamma^{t-s}|\mu_k(s) - \mu_k(t)|\mathbb{1}\{\pi = t\} \leq \gamma^{R(\gamma)}N_k(t - R(\gamma), \gamma). \tag{3.24}$$

As $N_k(t-R(\gamma),\gamma) \leq (1-\gamma)^{-1}$ we obtain $|V_k(t,\gamma)/N_k(t,\gamma)-\mu_k(t)|\leq \gamma^{R(\gamma)}((1-\gamma)N_k(t,\gamma))^{-1}$. Putting everything together we get

$$\left| \frac{V_k(t,\gamma)}{N_k(t,\gamma)} - \mu_k(t) \right| \leq \min\left\{1, \frac{\gamma^{R(\gamma)}}{(1-\gamma)N_k(t)}\right\}. \tag{3.25}$$

We use here the elementary inequality $\min\{1, x\} \leq \sqrt{x}$ and the way we defined $R(\gamma)$ to state that, for $t \in J(\gamma)$,

$$\left| \frac{V_k(t,\gamma)}{N_k(t,\gamma)} - \mu_k(t) \right| \leq \sqrt{\frac{\gamma^{R(\gamma)}}{(1-\gamma)N_k(t)}} \leq \sqrt{\frac{\xi \ln(n_K(\gamma))}{)N_k(t)}} \leq \frac{1}{2}c_k(t,\gamma) \tag{3.26}$$

which basically means that $R(\gamma)$ rounds after a change in distribution the bias can be bounded by half of the exploration bonus. We now take care of the fluctuations of the expected rewards $\hat{x}_k(t,\gamma)$ around $\frac{V_k(t,\gamma)}{N_k(t,\gamma)}$. If $t \in J(\gamma)$

$$\mathbb{P}(\mathcal{E}_k(t,\gamma)) \leq \mathbb{P}\left(\hat{x}_k(t,\gamma) \geq \mu_k(t) + \sqrt{\frac{\xi \ln(n_K(\gamma))}{N_k(t)}} + \left| \frac{V_k(t,\gamma)}{N_k(t,\gamma)} - \mu_k(t) \right| \right) \tag{3.27}$$

$$\leq \mathbb{P}\left(\hat{x}_k(t,\gamma) - \frac{V_k(t,\gamma)}{N_k(t,\gamma)} \geq \sqrt{\frac{\xi \ln(n_K(\gamma))}{N_k(t,\gamma)}}\right). \tag{3.28}$$

Step 4: We can define the total reward obtained by playing a determined arm k by

$$S_k(t) = N_k(t,\gamma)\hat{x}_k(t,\gamma). \tag{3.29}$$

We can use Theorem 3.2.3 and that $N_k(t,\gamma) \geq N_k(t,\gamma^2)$ because $\gamma > \gamma^2$ to state

$$\mathbb{P}(\mathcal{E}_k(t,\gamma)) \leq \mathbb{P}\left(\frac{S_k(t,\gamma) - V_k(t,\gamma)}{N_k(t,\gamma^2)} \geq \sqrt{\frac{\xi N_k(t,\gamma)\ln(n_K(\gamma))}{N_k(t,\gamma^2)}}\right) \tag{3.30}$$

$$\leq \mathbb{P}\left(\frac{S_k(t,\gamma) - V_k(t,\gamma)}{N_k(t,\gamma^2)} \geq \sqrt{\xi \ln(n_K(\gamma))}\right) \tag{3.31}$$

$$\leq \left\lceil \frac{\ln(n_t(\gamma))}{\ln(1+\eta)} \right\rceil n_t(\gamma)^{-2\xi(1-\eta^2/16)}. \tag{3.32}$$

We now substitute everything in Equation (135) while considering the expectation. For $\eta > 0$,

$$\mathbb{E}_\gamma[\overline{N}_k(T)] \leq 1 + \lceil T(1-\gamma)\rceil Q(\gamma)\gamma^{-1/(1-\gamma)} + R(\gamma)M_T \tag{3.33}$$

$$+ 2\sum_{t\in J(\gamma)} \left\lceil \frac{\ln(n_t(\gamma))}{\ln(1+\eta)} \right\rceil n_t(\gamma)^{-2\xi(1-\eta^2/16)}. \tag{3.34}$$

Recall that $\xi > \frac{1}{2}$ and thus $\eta = 4\sqrt{1-1/2\xi}$ so that $2\xi(1-\eta^2/16) = 1$. Given that, if we let $\tau = (1-\gamma)^{-1}$,

$$\sum_{t\in J(\gamma)} \left\lceil \frac{\ln(n_t(\gamma))}{\ln(1+\eta)} \right\rceil n_t(\gamma)^{-2\xi(1-\eta^2/16)} \leq \tau - K + \sum_{t=\tau}^{T} \left\lceil \frac{\ln(n_t(\gamma))}{\ln(1+\eta)} \right\rceil n_t(\gamma)^{-1} \tag{3.35}$$

$$\leq \tau - K + \left\lceil \frac{\ln(n_t(\gamma))}{\ln(1+\eta)} \right\rceil \frac{T}{n_k(t)} \tag{3.36}$$

$$\leq \tau - K + \left\lceil \frac{\ln\left(\frac{1}{1-\gamma}\right)}{\ln(1+\eta)} \right\rceil \frac{T(1-\gamma)}{1-\gamma^{1/(1-\gamma)}} \tag{3.37}$$

which concludes the proof. □

Theorem 3.2.1 shows that the expected regret yielded by the D-UCB algorithm is linear whereas our goal was to get a policy with sublinear regret. However, we can notice that the linear part of the expected regret depends on the input parameter of the algorithm γ. Thus, one choice might be to tune γ as a function of T. Specifically, if we let $\gamma = 1 - \frac{1}{4}\sqrt{\frac{M_T}{T}}$ then we obtain $\mathbb{E}[\overline{N}_k(T)] = O(\sqrt{TM_T}\ln(T))$.
Note that the number of change-points is arbitrary and might also be defined as a function of the time (i.e. $M_T = O(T^\alpha)$ with $\alpha \in [0,1)$). Therefore, if we tune parameter γ this way we obtain an expected regret which is sublinear of order $O(T^{\frac{1+\alpha}{2}}\ln(T))$. If we consider $\alpha = 0$ then the number of changes M is a constant and thus is independent of the time horizon T. Conceptually, it is normal that the more changes happen during the time interval, the more this strategy struggles to yield unbiased estimates of the expected rewards and thus the more it chooses a suboptimal arm. Similar things can be said about the SW-UCB algorithm, where if the window size is chosen to be $\tau = 2\sqrt{\frac{T\ln(T)}{M_T}}$ the expected number of times a suboptimal arm is played is $\mathbb{E}\left[\overline{N}_T(k)\right] = O(\sqrt{M_T T\ln(T)})$. Again, we can consider the number of changes to be a function of time so that $M_T = O(T^\alpha)$ with $\alpha \in [0,1)$ and then rewrite $\mathbb{E}\left[\overline{N}_T(k)\right] = O\left(T^{\frac{1+\alpha}{2}}\sqrt{\ln(T)}\right)$.
We can observe that whereas the expected regret achieved by the UCB policy in the stochastic stationary setting was of order $\ln(T)$, in this non-stationary setting two UCB-like strategies only manage to achieve expected regrets of order $T^{\frac{1+\alpha}{2}}\ln(T)$ and $T^{\frac{1+\alpha}{2}}\sqrt{\ln(T)}$. This leads to a natural question: how much can we improve the performances of the strategies in such an environment?

3.3 Lower Bound on the Regret of the Piecewise-Stationary Setting

We now want to discuss how much we can improve the performances in the abruptly changing environment and, thus, if we are able to identify a lower bound on the regret of any policy in this specific setting.
Let $N_k(s:t) = \sum_{u=s}^{t} \mathbb{1}\{\pi_u = i\}$ denote the number of times arm k is played between time s and t; Recall that $\sigma(X_{\pi_1}(1), ..., X_{\pi_t}(t))$ is the sigma algebra generated by the whole collection of rewards drawn before time t. We start by assuming that the rewards of each arm $k \in \{1, ..., K\}$ are generated by some probability distribution $P^{(k)}$ which does not change through time and let $\mathbb{P}_\pi$ be the rewards distribution yielded by the marginals on each arm P_k and the policy used π defined as

$$d\mathbb{P}_\pi(X(1:t)) = \prod_{s=1}^{t} dP^{(\pi_s)}(X_{\pi_s}(s)). \tag{3.38}$$

For ease of presentation assume that $\mu_1 > \mu_k \forall k \in \{1, ..., K\}$ We are going to divide the period $\{1, ..., T\}$ into epochs of the same size $\tau \in \{1, ..., T\}$ and change the distribution that generates

the payoffs in order to have arm K as optimal on one of these periods.
To do so, we define a distribution of rewards Q with expectation $\mu' > \mu(1)$, let $\delta = \mu' - \mu(1)$ and $\varrho = D(P_K; Q)$ be the Kullback-Leibler divergence between P_K and Q.
Let $\mathbb{P}_\pi^j$ be the modification of $\mathbb{P}_\pi$ where the j-th period of length τ has mean of the distribution generating rewards for arm K equal to μ'.
Denote by $\mathbb{E}_\pi^j[W]$ the expectation of the r.v. W under $\mathbb{P}_\pi^j$. Define as $\tilde{\mathbb{P}}_\pi$ the distribution of rewards when k is chosen uniformly at random in the set $\{1, ..., M\}$ and the expectation of a r.v. W under it as $\tilde{\mathbb{E}}_\pi[W] = M^{-1}\sum_{j=1}^M \mathbb{E}_\pi^j[W]$. Let $N^j(k) = N_k(1 + (j-1)\tau : j\tau)$ be the number of times arm k has been played in the $j - th$ epoch.

Theorem 3.3.1 (Garivier and Moulines [2011]). *For any horizon T such that $64/(9\varrho) \leq \mathbb{E}_\pi[N_K(T)] \leq T/(4\varrho)$ and for any policy π,*

$$\tilde{\mathbb{E}}_\pi[R_T] \geq C\frac{T}{\mathbb{E}_\pi[R_T]},$$

where $C = 2\delta(\mu_1 - \mu_K)/(3\varrho)$.

Proof. First, thank to the additivity of the K-L Divergence we can rewrite

$$D(\mathbb{P}_\pi||\mathbb{P}_\pi^j) = \sum_{t=1}^T D(\mathbb{P}_\pi(X_t|X_{1:t-1})||\mathbb{P}_\pi^j(X_t|X_{1:t-1})) \tag{3.39}$$

$$= \sum_{t=1+(j-1)\tau}^{j\tau} \mathbb{P}_\pi(\pi_t = K)D(P_K||Q) = \varrho\mathbb{E}_\pi[N_K(1 + (j-1)\tau : j\tau)] \tag{3.40}$$

where the second equality follows from the definition we gave of the distribution $\mathbb{P}_\pi^j$ which differs from $\mathbb{P}_\pi$ just in the j-th epoch. The last one follows as the term $D(P_K||Q)$ is independent of t and thus can be taken out of the sum and the sum of probabilities can be seen as a sum of expectations of indicator functions.
But

$$\mathbb{E}_\pi^j[N^j(K)] - \mathbb{E}_\pi[N^j(K)] \leq \tau d_{TV}(\mathbb{P}_\pi^j, \mathbb{P}_\pi) \leq \tau\sqrt{D(\mathbb{P}_\pi; \mathbb{P}_\pi^j)/2}$$

by Pinsker's inequality, which implies

$$\mathbb{E}_\pi^j[N^j(K)] \leq \tau\sqrt{D(\mathbb{P}_\pi; \mathbb{P}_\pi^j)/2} + \mathbb{E}_\pi[N^j(K)].$$

It is straightforward to observe that $\sum_{j=1}^M N^j(K) \leq N_T(K)$, which leads to

$$\sum_{j=1}^M \mathbb{E}_\pi^j[N^j(K)] - \mathbb{E}_\pi[N_T(K)] \leq \tau\sum_{j=1}^M \sqrt{\frac{\varrho\mathbb{E}_\pi[N^j(K)]}{2}} \tag{3.41}$$

$$\leq \tau\sum_{j=1}^M \sqrt{\frac{\varrho M\mathbb{E}_\pi[N_T(K)]}{2}}. \tag{3.42}$$

If the sum over the epochs is bounded from a quantity, it follows that there must exists an epoch $1 \leq j \leq M$ such that,

$$\tilde{\mathbb{E}}_\pi[N^j(K)] = \frac{1}{M}\sum_{k=1}^K \mathbb{E}_\pi^k[N^j(K)] \tag{3.43}$$

$$\leq \frac{1}{M}\mathbb{E}_\pi[N_T(K)] + \frac{\tau}{M}\sqrt{\frac{\varrho}{2}M\mathbb{E}_\pi[N_T(K)]} \tag{3.44}$$

$$\leq \frac{\tau}{T-\tau}\mathbb{E}_\pi[N_T(K)] + \sqrt{\frac{\varrho}{2}\frac{\tau^3}{T-\tau}\mathbb{E}_\pi[N_T(K)]} \tag{3.45}$$

where the last inequality follows from

$$\frac{1}{M} = \left\lceil \frac{\tau}{T} \right\rceil \leq \frac{\tau}{T} \leq \frac{\tau}{T-\tau}.$$

Under $\tilde{\mathbb{P}}_\pi$, the expectation of the regret R_T can be lower-bounded as

$$\frac{\tilde{\mathbb{E}}_\pi[R_T]}{\delta} \geq \tau - \tilde{\mathbb{E}}_\pi[N_T(K)] \tag{3.46}$$

$$\geq \left(\tau - \frac{\tau}{T-\tau}\mathbb{E}_\pi[N_T(K)] + \sqrt{\frac{\alpha}{2}\frac{\tau^3}{T-\tau}\mathbb{E}_\pi[N_T(K)]} \right) \tag{3.47}$$

We can maximize the right hand side of equation (3.47) by choosing either $\tau = 8T/9\varrho\mathbb{E}_\pi[N_T(K)]$ or $M = 9\varrho/(8\mathbb{E}_\pi[N_T(K)])$. This leads to the lower-bound:

$$\frac{\tilde{\mathbb{E}}_\pi[R_T]}{\delta} \geq \frac{16\delta}{27\alpha}\left(1 - \frac{\mathbb{E}_\pi[N_T(K)]}{T}\right)^2 \left(1 - \frac{8}{9\alpha\mathbb{E}_\pi[N_T(K)]}\right)\frac{T}{\mathbb{E}_\pi[N_T(K)]}. \tag{3.48}$$

We conclude the proof by noting that $\mathbb{E}_\pi[N_T(K)] \leq \mathbb{E}[R_T]/(\mu_1 - \mu_K)$ and thus

$$\frac{\tilde{\mathbb{E}}_\pi[R_T]}{\delta} \geq \frac{16\delta(\mu_1-\mu_K)}{27\varrho}\left(1 - \frac{\mathbb{E}_\pi[N_T(K)]}{T}\right)^2 \left(1 - \frac{8}{9\varrho\mathbb{E}_\pi[N_T(K)]}\right)\frac{T}{\mathbb{E}_\pi[R_T]}. \tag{3.49}$$

□

In words, Theorem 3.3.1 states that whatever policy we are using to address a multi-armed bandit problem where at least two changes on the same arm happened, we are going to incur an expected regret $\tilde{\mathbb{E}}_\pi[R_T]$ which is for sure greater than a quantity which is a function of time and which is inversely proportional to the regret on a problem with same initial distribution means but no changes $\mathbb{E}_\pi[R_T]$. Another thing to note is that in the lower bound we have a linear factor T. However, the interpretation of this linear factor in the lower-bound is strictly related to the regret of the policy in the stationary case. Recall that for a lot of policies we have their upper-bound for the stationary case. Thus Theorem 3.3.1 can help us in observing that if a policy achieves an expected regret of order $O(\ln(T))$ in the stationary case, then its expected regret in the non-stationary abruptly changing environment is going to be lowerbounded as $\tilde{\mathbb{E}}_\pi[R_T] \geq C\frac{T}{\ln(T)}$. An usuful corollary of the theorem aboves states that in a non-stationary setting no policy can achieve an expected regret of order lower than $\sqrt{T}$.

Corollary 3.3.1 (Garivier and Moulines [2011]). *For any policy π and any positive time horizon T we have*

$$\max\{\mathbb{E}_\pi[R_T], \tilde{\mathbb{E}}_\pi[R_T]\} \geq \sqrt{CT}$$

Proof. If the condition $\mathbb{E}_\pi[N_K(T)] \leq 16/(9\varrho)$ or $\mathbb{E}_\pi[N_K(T)] \geq T/\varrho$ holds then the corollary obviously holds. Otherwise we have that, because of Theorem 10,

$$\max\{\mathbb{E}_\pi[R_T], \tilde{\mathbb{E}}_\pi[R_T]\} \geq \max\{\mathbb{E}_\pi[R_T], C\frac{T}{\mathbb{E}_\pi[R_T]}\} \geq \sqrt{CT}\} \tag{3.50}$$

□

Next, we are going to analyze how different policies behave in terms of expected regret for the piecewise-stationary with abrupt changes setting and how close they get to the provided lower bound. Different policies are going to have different features. As previously mentioned, D-UCB and SW-UCB both relied on penalizing old observation even if with different approaches. While we can notice that those policies were not aiming to spot the changes, one may wonder whether it might be worth it to do so and what benefits we might get from it.

3.4 Monitored-UCB

The next policy we introduced is called Monitored-UCB (M-UCB) (Cao et al. [2019]). As the mentioned at the end of the previous section, a interesting feature that a strategy might have in this setting is an integrated change-point detection algorithm to spot where the changes happen. The authors of this paper assumed the changepoints to always be global.
Change-point detection is a topic that has been studied deeply in the past decades and that has had huge impact on the study of change-point detection for time series. Many type of sofisticated algorithms have been designed, like CUSUM (PAGE [1954]) and the GLR (generalized likelihood ratio) procedure defined in Willsky and Jones [1976]. The above mentioned methods, however, usually assumes the parameters of the distributions known which is not necessarily the case in the setting we are considering.
Therefore, the change-point detection algorithm designed for the M-UCB strategy is a rather simple procedure which, given a sequence of realizations from independent random variables $X_1, ..., X_w$ with w even, it checks if a change happened in the middle of the sequence by comparing the first half of the sequence with the second half of it. If the difference between the sum of the first half and the second is bigger than a quantity b then it triggers an alarm of change in distribution. More specifically, the consistency of Algorithm 8 is given by MCDiarmid's inequal-

Algorithm 8 Change-Point Detection Procedure

Require: $b, w, X_1, ..., X_w$.
Ensure: Play each arm once.

1: **if** $|\sum_{t=1}^{w/2} X_t - \sum_{t=w/2+1}^{w} X_t| > b$ **then**
2: **Return True**
3: Else **Return False**

ity which allows to bound the probability of the algorithm returning true when there is no change with $2\exp(-2b^2/w)$ and the probability of returning false when there was a change at $w/2$ with $1-2\exp(wc^2/4)$ with $c = |\mathbb{E}[X_{w/2+1}] - \mathbb{E}[X_{w/2}]|$ being the index of the magnitude of the change between old and new mean.
The main idea behind the M-UCB policy is to perform in parallel the classic UCB policy together with sequential change-point detection on the arm played by the UCB policy. One might argue that the classic UCB policies were shown to be suboptimal for this setting. However, the strategy is based on the concept of *restarting the system completely* whenever a change on one arm is spotted. Thus, we can think of this policy as a fragmentation of the timeline in multiple shorter stationary problems.
This brings us to introdue the other main idea behind this algorithm. Given that one of the goals is to spot changes, we want to be sure that we do not miss changes on the totality of the arms. Thus, we introduce the concept of uniform sampling coefficient γ (that has nothing to do with the discount factor of the previous section). For a fraction γ of the time we are going to sample deterministically from all arms in order to have some samples from suboptimal arms as well and check for changes. Let also τ be the time of the last detection, again this notation must not be confused with the window size τ from the SW-UCB policy. Denote with n_k the number of times arm k has been played after time τ.
A further assumption made is that once a change happens, the next one will happen at least L time steps later. These L steps are used by the algorithm to collect information from the new distribution. As mentioned above, this policy has a non zero probability of missing a change in distribution which is linked to the magnitude of each change. We are therefore interested in the magnitude of the changes of each arm. Therefore, let the amplitude of the change of arm k at the i-breakpoint be denoted as

$$\Delta_k^{(i)} = |\mu_k^{(i+1)} - \mu_k^{(i)}| \tag{3.51}$$

Algorithm 9 Monitored UCB. [Cao et al., 2019]

Require: T, K, number of cpts m_T,b, L, an even integer $w > 0$.

Ensure: Play each arm once and update the empirical mean for each arm, set $\gamma = \frac{w \cdot K}{L}$, $\tau \leftarrow 0$ and $n_k \leftarrow 0 \; \forall k \in \{1, ..., K\}$

Phase 1 - Choice between uniform and optimal arm sampling

1: **for** $t = 1, 2, ..., T$ **do**
2: $A \leftarrow (t - \tau) \bmod \left\lceil \frac{K}{\gamma} \right\rceil$
3: **if** $A \leq K$ **then**
4: $\pi_t \leftarrow A$ Line 2/5 decide whether to perform uniform sapmling or not
5: **else**
6: **Arm Selection:**
7: **for** $k = 1, ..., K$ **do**

$$MUCB_k \leftarrow \frac{1}{n_k} \sum_{t=1}^{n_k} Z_k(n) + \frac{2 \log(t - \tau)}{n_k}$$

end for
8: $\pi_t \leftarrow \text{argmax}_{k \in \{1,...,K\}} MUCB_k$
9: **end if**
10: $n_{\pi_t} \leftarrow n_{\pi_t} + 1$
11: $Z_{\pi_t}(n_{\pi_t,t}) \leftarrow X_{\pi_t}(t)$

Phase 2 - Check for changes in distribution

12: **if** $(t - \tau) > L, \; n_{\pi_t} > w$ **then**
13: $CD_t \leftarrow |\sum_{i=1}^{w/2} Z_{\pi_t}(n_{\pi_t,t-i}) - \sum_{i=w/2+1}^{w} Z_{\pi_t}(n_{\pi_t,t-i})|$
14: **if** $CD_t > b$ **then**
15: $\tau \leftarrow t$
16: $n_k \leftarrow 0 \; \forall k \in \{1, ..., K\}$
17: **end if**
18: **end if**
19: **end for**

and the maximum change happened at the i-th changing as

$$\Delta^{(i)} = \max_{k \in \{1,..,K\}} \Delta_k^{(i)}. \tag{3.52}$$

Let us also define a quantity $\varsigma_k^{(i)}$ referring to how much suboptimal was arm k on the i-th segment as

$$\varsigma_k^{(i)} = \max_{k' \in \{1,...,K\}} \{\mu_{k'}^i\} - \mu_k^i. \tag{3.53}$$

We mentioned previously that this policy requires each segment to be long at least L to obtain informations about the new distribution on all the arms. Therefore, if we require at least w samples before the strategy starts looking for changes and if we perform uniform sampling with frequency γ we can set the minimum length of each segment to be $L = w \left\lceil \frac{K}{\gamma} \right\rceil$. This guarantees that when the distributions can change again we will have enough samples from each arm to detect it. We now formalise the assumptions we declared previously and add one more.

Assumption 3.4.1 ([Cao et al., 2019]). *Let w and γ be two parameters chosen from the decision maker so that $m_T < \lceil T/L \rceil - 1$, $\nu_{i+1} - \nu_i > L$ for each $i \in \{1, ..., m_T + 1\}$. Then we assume that there exists a $k \in \{1, ..., K\}$ so that for each $i \in \{1, ..., m_T + 1\}$ it holds*

$$\Delta_k^{(i)} \geq 2\sqrt{\ln(2KT^2)/w} + 2\sqrt{\ln(2T)/w}. \tag{3.54}$$

The last assumption we make is that at each break-point there is at least an arm k with a change of magnitude big enough to be able to detected by the algorithm. Without this assumption we could risk to miss the changes on all the arms despite of the samples from each arm as the magnitude of the changes would be too small to allow the detection trigger to be activated.
We now proceed to show the bound the expected regret of the M-UCB policy using the defined tools.

Theorem 3.4.1 ([Cao et al., 2019]). *The expected regret incurred by M-UCB policy when γ and w satisfy Assumption 1 and $b = \sqrt{w \ln(2KT^2)/2}$ is bounded as*

$$\mathbb{E}_\pi[R(T)] \leq \sum_{i=1}^{m_T+1} C_i + \gamma T + \sum_{i=1}^{m_T} \frac{K \min\{\frac{w}{2}, \left\lceil \frac{b}{\Delta^{(i)}} \right\rceil + 3\sqrt{w}\}}{\gamma} + 3(m_T + 1) \tag{3.55}$$

where $C_i = 8 \sum_{i:\varsigma_k^{(i)}>0} \frac{\ln(T)}{\varsigma_k^{(i)}} + \left(1 + \frac{\pi^2}{3} + K\right) \sum_{i:\varsigma_k^{(i)}>0} \varsigma_k^{(i)}$

The expected regret above is articulated in four different terms where each of these represents the expected regret coming from different aspects of the strategy. The first term represent the expected regret coming from the UCB exploration. More specifically, this part of the regret is strongly dependent on the quantities $\varsigma_k^{(i)}$. The second term appearing in the expected regret expression displays the regret yielded by the uniform sampling component of the strategy, as for a fraction γ of the time T the policy will sample suboptimal arms it can be bounded by γT. The last two terms come from the change-point detection algorithm in the strategy: while the third term derives from the expected delay of detection the last one bounds the component incurred when a break-point is missed or when we have a false detection.
Before moving into the details of the regret analysis we show how the policy performs if the parameters are tuned optimally.

Corollary 3.4.1 ([Cao et al., 2019]). *The expected regret incurred by M-UCB policy with $w, b = \sqrt{w \ln(2KT^2)/2}$ and*

$$\gamma = \sqrt{\sum_{i=1}^{m_T} K \min\{\frac{w}{2}, \left\lceil \frac{b}{\Delta^{(i)}} \right\rceil + 3\sqrt{w}\}/(2T)} \tag{3.56}$$

can be bounded as

$$\mathbb{E}_\pi[R(T)] \leq \sum_{i=1}^{m_T+1} C_i + \gamma T + \sqrt{\sum_{i=1}^{m_T} 2TK \min\{\frac{w}{2}, \left\lceil \frac{b}{\Delta^{(i)}} \right\rceil + 3\sqrt{w}\}} + 3(m_T - 1) \tag{3.57}$$

where $C_i = 8 \sum_{i:\varsigma_k^{(i)}>0} \frac{\ln(T)}{\varsigma_k^{(i)}} + \left(1 + \frac{\pi^2}{3} + K\right) \sum_{i:\varsigma_k^{(i)}>0} \varsigma_k^{(i)}$

Following the tuning we provided above we can also define an optimal tuning for the input parameter w.

Remark 3.4.1 ([Cao et al., 2019]). *Let the minimum length of the segment between two consecutive break-points L be big enough. Then we can set the minimum sample size w as*

$$w = (4/\Delta^2)\left(\sqrt{\ln(2KT^2)} + \sqrt{\ln(2T)}\right)^2 \tag{3.58}$$

to satisfy Assumption 1.

We now introduce a series of lemmas that will be helpful in the regret analysis of the M-UCB strategy. The first lemma we show deals with bounding the expected regret of the policy in a stationary scenario (i.e. $m_T = 1$).

Lemma 3.4.1 (Regret bound for the M-UCB in stationary scenarios [Cao et al., 2019])**.** *Consider a stationary scenario with $m_T = 1,\ \nu_0 = 0, \nu_1 = T$. Under Algorithm 1 with parameter w, b, γ we have that*

$$\mathbb{E}[R(T)] \leq T \cdot P(\tau_1 \leq T) + \tilde{C} + \gamma T$$

where τ_1 is the first detection time and

$$\tilde{C} = 8 \sum_{\varsigma_k^{(i)}} \frac{\log(T)}{\varsigma_k^{(i)}} + (1 + \frac{\pi^2}{3} + K) \sum_{k=1}^{K} \varsigma_k^{(i)}$$

Proof. We begin the proof by providing a decomposition of the expected regret expression.

$$\mathbb{E}[R(T)] = \mathbb{E}[R(T)\mathbb{1}\{\tau_1 \leq T\}] + \mathbb{E}[R(T)\mathbb{1}\{\tau_1 > T\}] \tag{3.59}$$
$$\leq T \cdot \mathbb{P}(\tau_1 \leq T) + \mathbb{E}[R(T)\mathbb{1}\{\tau_1 > T\}] \tag{3.60}$$

where the inequality follows from bounding $\mathbb{E}[R(T)]$ with T as $R(T) \leq T$ by definition. Recall that $N_k(t)$ represents the number of times arm k has been played by time t. The second term in equation (3.59) is the expected regret incurred when there is no detection. As this lemma referes to the stationary case if no changes are spotted it means that the change-point detection algorithm did not perform any mistake and unnecessary restart. We can rewrite

$$\mathbb{E}[R(T)\mathbb{1}\{\tau_1 > T\}] = \sum_{\varsigma_k}^{(1)} \mathbb{E}[N_k(t)\mathbb{1}\{\tau_1 > T\}]. \tag{3.61}$$

Thus, it remain to upperbound the expected number of times a suboptimal arm has been played given that no changes have been spotted $\mathbb{E}[N_k(T)\mathbb{1}\{\tau_1 > T\}]$.
We start by considering $N_k(T)\mathbb{1}\{\tau_1 > T\}$.

$$N_k(T)\mathbb{1}\{\tau_1 > T\} = \sum_{t=1}^{T} \mathbb{1}\{\pi_t = k, \tau_1 > T\} \tag{3.62}$$
$$= \sum_{t=1}^{T} \mathbb{1}\{\pi_t = k, \tau_1 > T, N_k(t) < l\} + \sum_{t=1}^{T} \mathbb{1}\{\pi_t = k, \tau_1 > T, N_k(t) \geq l\} \tag{3.63}$$
$$\leq l + \sum_{t=1}^{T} \mathbb{1}\{\pi_t = k, \tau_1 > T, N_k(t) \geq l\} \tag{3.64}$$
$$\leq l + \sum_{t=1}^{T} \mathbb{1}\{t \bmod \left\lceil \frac{K}{\gamma} \right\rceil = k, N_k(t) \geq l\} \tag{3.65}$$
$$+ \sum_{t=1}^{T} \mathbb{1}\{k = \underset{k' \in \{1,\dots,K\}}{\operatorname{argmax}} MUCB_{k'}, N_k(t) \geq l\} \tag{3.66}$$
$$\leq l + \left\lceil \frac{T\gamma}{K} \right\rceil + \sum_{t=1}^{T} \mathbb{1}\{k = \underset{k' \in \{1,\dots,K\}}{\operatorname{argmax}} MUCB_{k'}, N_k(t) \geq l\} \tag{3.67}$$

where the first inequality follow from bounding a sum of at maximum l non-zero indicator functions with l. The second comes from the fact that if an arm has been sampled at time t it means

that that arm must have been either optimal at that time or uniform sampled.
If we set $l = \lceil 8 \ln T / \varsigma_k^{(1)} \rceil$, using the same approach we showed in Theorem 3.4.1 we obtain

$$\mathbb{E}[N_k(t)\mathbb{1}\{\tau_1 > T\}] \leq \frac{T\gamma}{K} + \frac{8\ln(T)}{(\varsigma_k^{(1)})^2} + 1 + \frac{\pi^2}{3} + K. \tag{3.68}$$

Summing over k yields the desired result. □

We can observe that in a context where no changes can occur the expected regret is simply reduced to the terms regarding the exploration coming from both the uniform sampling and the UCB procedures. The first term in the expression deals with the expected regret yielded if the system spots a change where no changes happen.

We now want to bound the probability still to be bounded in Lemma 3.4.1. More specifically we want to bound the probability of raising a false alarm.

Lemma 3.4.2 (Probability of raising one false alarm [Cao et al., 2019])**.** *Consider a stationary scenario with $m_T = 0$. Then under Algorithm 9 with parameter $w < T, b$ and γ, we have that*

$$P(\tau_1 \leq T) \leq wK \left(1 - \left(1 - 2\exp(-\frac{2b^2}{w})\right)^{\lceil \frac{T}{w} \rceil}\right)$$

where τ_1 is the first detection time.

Proof. Let us introduce the time of the first detection on the k-th arm $\tau_{k,1}$. The time of first detection is then written as $\tau_1 = \min_{k \leq K}\{\tau_{k,1}\}$. Thank to the union bound we can write

$$P(\tau_1 < T) \leq \sum_{k=1}^{K} P(\tau_{k,1} < T). \tag{3.69}$$

Given this and recalling that w is the minimum sample size for each arm to start looking for changes, we can define the quantity $S_{k,t}$ for each arm k and $j \geq w$ as

$$S_{k,t} = \left| \sum_{i=t-w/2+1}^{t} Z_k(i) - \sum_{i=t-w+1}^{t-w/2} Z_k(i) \right| \tag{3.70}$$

We can then define more formally the first time of detection on arm k as

$$\tau_{k,1} = \inf\{t \geq w : S_{k,t} > b\}. \tag{3.71}$$

We now define, for each $0 < j \leq w - 1$,

$$\tau_{k,1}^{(j)} = \inf = \{t = j + nw, n \in \mathbb{N}, S_{k,t} > b\}. \tag{3.72}$$

It follows directly that $\tau_{k,1} = \min\{\tau_{k,1}^{(1)}, ..., \tau_{k,1}^{(w-1)}\}$. We can observe that if the environment is stationary then the distribution, for each $0 < j \leq w - 1$, is a geometric distribution:

$$\mathbb{P}(\tau_{k,1}^{(j)} = nw + j) \leq p(1-p)^{n-1} \tag{3.73}$$

where $p = \mathbb{P}(S_{k,t} > b)$. Therefore, we use again the union bound to write

$$\mathbb{P}(\tau_{k,1}) \leq w\left(1 - (1-p)^{\lceil \frac{T}{w} \rceil}\right). \tag{3.74}$$

Last, we have to bound the probability that the difference between the two halves of a segment of size w is greater than b when the distribution is stationary. To do so we make use of the McDiarmid's inequality to write that

$$p \leq 2\exp\left(-\frac{2b^2}{w}\right). \tag{3.75}$$

□

With this lemma and its proof we have therefore put a bound on the expected regret that we can incur when using the M-UCB policy on a stationary scenario and therefore on the false detections and suboptimal plays rising from UCB nature.
The result we just saw above yields to the optimal tuning of the input parameter b.

Remark 3.4.2. *Result from Lemma 3.4.2 implies that under the above mentioned conditions we have* $P(\tau_1 \leq T) \leq wK\left(1-\left(1-2\exp(-\frac{2b^2}{w})\right)^{\lceil \frac{T}{w} \rceil}\right)$. *Thus if we set* $b = \sqrt{\frac{w\ln(2KT^2)}{2}}$ *we obtain* $P(\tau_1 \leq T) \leq 1/T$ *which means that in a game of length* T *the policy is expected to raise an alarm when not needed at most once per game.*

We now move to analyze how the algorithm behaves when it comes to detecting actual changes and what is the probability of achieving a successful detection.

Lemma 3.4.3 (Probability of achieving a successful detection [Cao et al., 2019])**.** *Consider a piecewise-stationary scenario with* $m_T = 1$ *and recall that* $L = w\left\lceil \frac{K}{\gamma} \right\rceil$. *Recall that* $\nu_1 - \nu_0 > L$. *For any* $\mu^1, \mu^2 \in [0,1]^K$ *satisfying* $\delta_{\tilde{k}} \geq 2\frac{b}{w} + c$ *for some* $\tilde{k} \leq K$ *and* $c > 0$, *under Algorithm 9, we have that*

$$P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1\right) \geq \left(1 - 2\exp\left(-\frac{wc^2}{4}\right)\right)$$

Proof.

$$\begin{aligned}
&P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1\right) \\
&\geq P(S_{\tilde{k},w} > b) \\
&\geq \left(1 - 2\exp\left(-\frac{(w\Delta_{\tilde{k}}^{(1)}/2 - b)^2}{4}\right)\right) \\
&\geq \left(1 - 2\exp\left(-\frac{wc^2}{4}\right)\right)
\end{aligned}$$

where the inequalities follow from MCDiarmid's inequality and Assumption 3.4.1. □

Before moving forward we propose another remark on the optimal tuning of the input parameters.

Remark 3.4.3 ([Cao et al., 2019])**.** *If we let* $b = \sqrt{\frac{w\ln(2KT^2)}{2}}$ *and* $c = \sqrt{\frac{\ln(2T)}{w}}$ *then we can lower-bound the probability of achieving a successful detection with* $1 - 1/T$. *Similarly as the previous remark, it means that in a game of length* T, *the policy is expected to fail to raise an alarm when needed at most once per game.*

The last lemma we introduce before moving into the actual proof of the regret analysis manages to find a bound on the expected delay of detection yielded by the M-UCB policy.

Lemma 3.4.4 (Expected detection delay Cao et al. [2019])**.** *Consider a piecewise-stationary scenario with $m_T = 2$, and recall that $L = w\left\lceil \frac{K}{\gamma} \right\rceil$. For any $\mu^1, \mu^2 \in [0,1]^K$ satysfing $\delta_{\tilde{k}}^{(1)} > 2\frac{b}{w} + c$ for some $\tilde{k} \leq K$, we have that*

$$E\left[\tau_1 - \nu_1 \mid \nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2}\right] \leq \frac{\min(\frac{L}{2}, \left\lceil b/\delta_{\tilde{k}}^{(1)} \right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil)}{1 - 2\exp(-w\frac{c^2}{4})}$$

Proof. We define $N = \left\lceil \frac{b}{\Delta_{\tilde{k}}^{(1)}} \right\rceil \cdot \left\lceil \frac{K}{\gamma} \right\rceil$ and we use it to upper bound the Expected Detection Delay as

$$\begin{aligned} E\left[\tau_1 - \nu_1 \mid \nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2}\right] &= \sum_{i=1}^{TL/2} P\left(\tau_1 \geq \nu_1 + i \mid \nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2}\right) \\ &\leq N + \sum_{i=N}^{TL/2} P\left(\tau_1 \geq \nu_1 + i \mid \nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2}\right). \end{aligned}$$

Because of the the choice of N we are guaranteed to have at least $\frac{i}{\lceil K/\gamma \rceil}$ samples from each arm within i time steps. We can use McDiarmid's inequality and Lemma 3.4.3 to have that

$$\begin{aligned} &\sum_{i=N}^{TL/2} P(\tau \geq \nu_1 + i \mid \nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2}) \\ &= \sum_{i=N}^{TL/2} \frac{P(\nu_1 + i \leq \tau_1 \leq \nu_1 + TL/2 \mid \tau_1 > \nu_1)}{P(\nu_1 \leq \tau_1 \leq \ \nu_1 + TL/2 \mid \tau_1 > \nu_1)} \\ &\leq \frac{1}{1 - 2\exp(-wc^2/4)} \sum_{i=N}^{TL/2} 2\exp\left(-\frac{(i/\lceil (K/\gamma) \rceil\, \Delta_{\tilde{k}}^{(1)} - b)^2}{w}\right) \\ &\leq \frac{(K/\gamma)}{1 - 2\exp(-wc^2/4)} \sum_{j=\left\lceil b/\Delta_{\tilde{k}}^{(1)} \right\rceil}^{w/2} 2\exp\left(-\frac{(j\Delta_{\tilde{k}}^{(1)} - b)^2}{w}\right). \end{aligned}$$

Define $q = \left\lceil (w/2) \cdot \Delta_{\tilde{k}}^{(1)} \right\rceil - b$ and we have $q > 1$ from the assumption that $\Delta_{\tilde{k}}^{(1)} > 2b/w + c$. Thus, we have

$$\begin{aligned} &(1 - 2\exp(-wc^2/4)) \cdot E\left[\tau_1 - \nu_1 \mid \nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2}\right] \\ &\leq N + \left\lceil \frac{K}{\gamma} \right\rceil \cdot \sum_{j=\left\lceil b/\delta_{\tilde{k}}^{(1)} \right\rceil}^{w/2} 2\exp\left(-\frac{(j\delta_{\tilde{k}}^{(1)} - b)^2}{w}\right) \\ &\leq N + 2\left\lceil \frac{K}{\gamma} \right\rceil \cdot \left(1 + \int_1^q \exp\left(-\frac{l^2}{w}\right) dl\right) \\ &\leq N + 2\left\lceil \frac{K}{\gamma} \right\rceil \cdot \left[1 + \sqrt{w}\left(1 - \frac{1}{\sqrt{w}} + \int_1^{q/\sqrt{w}} \exp(-u^2) du\right)\right] \\ &\leq N + 2\left\lceil \frac{K}{\gamma} \right\rceil \cdot \left(\sqrt{w} + \sqrt{w}\int_1^{q/\sqrt{w}} u\exp(-u^2) du\right) \\ &\leq (\left\lceil b/\Delta_{\tilde{k}}^{(1)} \right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil. \end{aligned}$$

The third inequality follows from a change of variable (i.e. l into $u = \frac{l}{\sqrt{w}}$) and the fourth one from the fact that $\exp(-u^2) \leq u\exp(-u^2), u \geq 1$ in the first inequality. By the definition of the conditioning event we also have that

$$(1 - 2\exp(-wc^2/4)) \cdot E\left[\tau_1 - \nu_1 \mid \nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2}\right]$$

Combining the above analysis we conclude the result. □

3.4.1 Proof of Theorem 3.4.1

Now, given the four lemmas that we proposed above, we show the proof of the regret analysis.

Proof. Recall that $L = w\left\lceil \frac{K}{\gamma} \right\rceil$. Define events $F_i = \{\tau_i > \nu_i\}$ and $W_i = \{\tau_i < \nu_i + \frac{TL}{2}\}$. It follows that the event $F_i \cap W_i$ is the event where the $i+1$-th change is not too close to the i-th one and it can be detected correctly and efficiently.
Define $R(T) = \sum_{t=1}^{T} \max_{k \in K} X_{k,t} - X_{A_t,t}$ and thus $\mathcal{R}(T) = E[R(T)]$. Thus

$$\begin{aligned}
\mathcal{R}(T) = E[R(T)] &\leq E[R(T)\mathbb{1}\{F_1\}] + E[R(T)\mathbb{1}\{\overline{F}_1\}] \\
&\leq E[R(T)\mathbb{1}\{F_1\}] + T \cdot (1 - P(\mathcal{F}_i)) \\
&\leq E[R(T)\mathbb{1}\{F_1\}] + 1 \\
&\leq E[R(\nu_1)\mathbb{1}\{F_1\}] + E[R(T) - R(\nu_1)] + 1 \\
&\leq C_1 + \gamma\nu_i + E[R(T) - R(\nu_i)] + 1
\end{aligned}$$

where the second inequality follows from the classic upper-bounding, the third one follows from Lemma 3.4.2 by setting $b = \sqrt{w\log(2KT^2)/2}$ and the last one follows from Lemma 3.4.1 (the segment of time before ν_1, time of the first change, is clearly stationary). Therefore, we have found an upper-bound on the first segment of time.
We have now to upper-bound $E[R(T) - R(\nu_i)]$.

$$\begin{aligned}
E[R(T) - R(\nu_i)] &\leq E[R(T) - R(\nu_1) \mid F_1, W_1] + T \cdot (1 - P(F_1, W_1)) \\
&\leq E[R(T) - R(\nu_1) \mid F_1, W_1] + 2
\end{aligned}$$

where the inequality follows from Lemma 3.4.3.
Now, we only have to bound the term of the regret conditioned on the "good events" to happen.

$$\begin{aligned}
E[R(T) - R(\nu_1)|F_1, W_1] &\leq E[R(T) - R(\tau_1)|F_1, W_1] + E[R(\tau_1) - R(\nu_1)|F_1, W_1] \\
&\leq \tilde{E}[R(T - \nu_1)] + E[\tau_1 - \nu_1|F_1, W_1] \\
&\leq \tilde{E}[R(T - \nu_1)] + \min\left(\frac{TL}{2}, (\left\lceil b/\Delta_{\tilde{k}}^{(1)} \right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil\right)
\end{aligned}$$

where $\tilde{E}$ is the expectation according to the bandit problem starting from the second segment. The second inequality follows from the renewal property of the algorithm (after spotting a change everything restarts) whereas the third comes from Lemma 3.4.4.
Thus,

$$E[R(T)] \leq \tilde{E}[R(T - \nu_1)] + \tilde{C}_1 + \gamma\nu_1 + \min\left(\frac{TL}{2}, (\left\lceil b/\Delta_{\tilde{k}}^{(1)} \right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil\right)$$

Thanks to the recursion we defined, we can conclude that

$$\begin{aligned}
E[R(T)] \leq &\sum_{i=1}^{m_T+1} C_i + \gamma T + \sum_{i=1}^{m_T} \min\left(\frac{TL}{2}, (\left\lceil b/\Delta_{\tilde{k}}^{(i)} \right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil\right) \\
&+ (m_T + 1)
\end{aligned}$$

□

The bound found on the expected regret is then of order $O(\sqrt{MKT\ln(T)})$. This means that the regret bound yielded on this policy is close to the lower bound on the expected regret we introduced with Corollary 3.4.1 up to a logarithmic factor. Therefore, if we have to compare the regret bound found here with the one yielded on the SW-UCB policy we can notice that they achieve the same bound for the part depending on the time horizon T. Therefore, introducing the change-point detector does not seem to have an impact of the bound.

Chapter 4

Experimental Results

In the previous section we explored some UCB type of policies for the non-stationary setting. We saw that while M-UCB was performing detection of the change-points, D-UCB was based on discounting old observations without any detection. However, as we pointed out, designing a policy with a change-point detector and a restarting procedure does not seem to tighten the expected regret of a UCB policy to its lower bound. Therefore, we might wonder if it is actually helpful to use a change-point detector in a policy using these settings.

M-UCB is in fact assuming the number of changes to be given in advance. Thus, we might argue that knowing the number of changes in advance might make change-point detection sub-optimal as detection involves the probability of not detecting a change or raising a false alarm which might lead to an extremely bad regret. D-UCB in this context seems to be more solid as it also does not require a minimum magnitude on at least one change per break-point time and it does not require a minimum number of time steps between a change and the next one.

However, it might be interesting to see how M-UCB would work by stretching these assumptions and also to see under which conditions using a change-point detector might be better.

4.1 Simulations

We start by showing how the expected error and regret behave for the M-UCB policy in the standard setting for the strategy. We are considering a three arms scenario with $T = 10000$ where on arm 1 we have $\mu_1^{(i)} = 0.9$ when i is odd and $\mu_1^{(i)} = 0.1$ when i is even and the opposite on arm 2. Arm 3 has constantly mean $\mu_3^{(i)} = 0.5$ for all $i \in \{1, ..., m_T + 1\}$. We assume to have three changes at times $2500, 5000$ and 7500.
Concerning the input parameters, w is set to be $w = 200$ whereas the other parameters are set accordingly to the optimal tuning proposed for the M-UCB strategy. Therefore we obtain the tuned parameters $b = \sqrt{\frac{200 \ln(2 \cdot 3 \cdot (10000)^2)}{2}} = \sqrt{100 \cdot 21.6} = 46.47$ and $\gamma = \sqrt{m_T K (2b + 2\sqrt{w})/2} = \sqrt{3 \cdot 3(2 \cdot 46.47 + 2\sqrt{200})/2} = 23.36$.

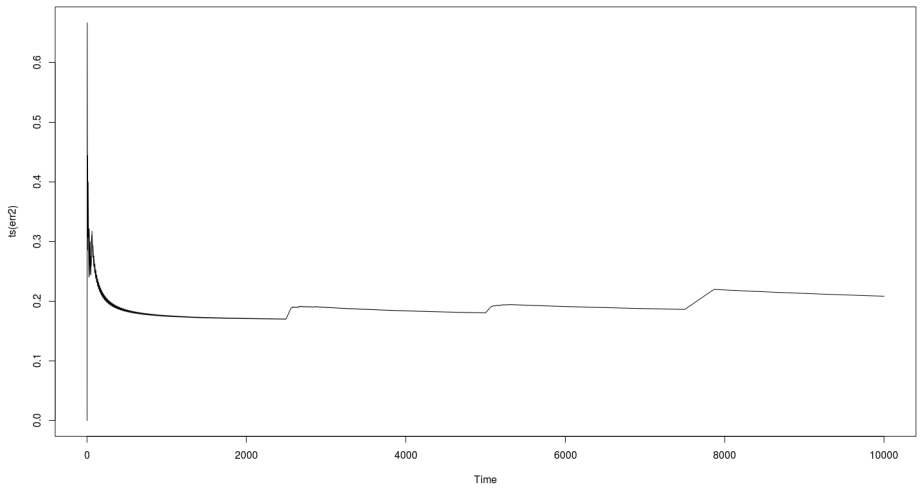

Figure 4.1: Error of the M-UCB policy in its standard setting

We can notice that in the beginning the error assumes a high value which decreases quickly. This is due to the lack of information in the early steps of the game and also because in the first K steps of the game we play each arm once. We can also observe that the regret decreases steadily until a change happens and we observe small drifts until the change is spotted and the system is restarted. The plot highlights that the drifts happen in time periods close to the real changes. Because of the way the algorithm is constructed, the policy works like a standard UCB in each stationary segment.

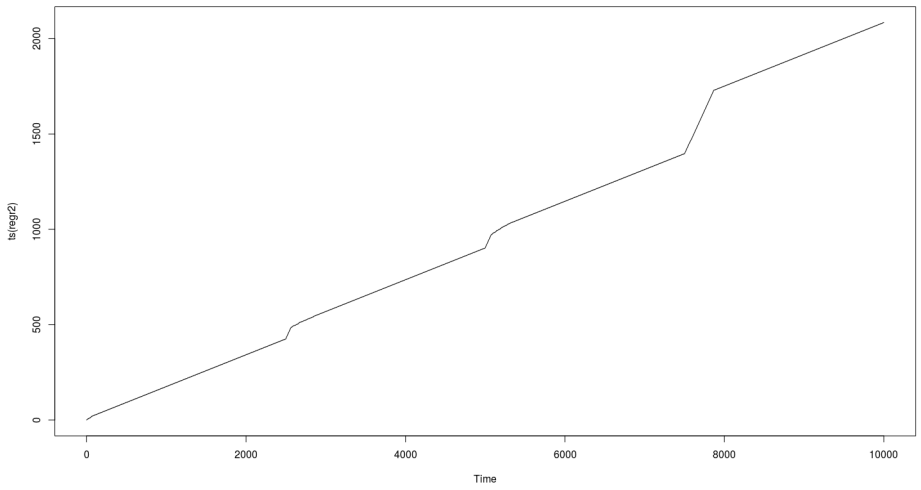

Figure 4.2: Regret of the M-UCB policy in its standard setting

The regret plot shows how the regret has a linear component which is given by the uniform sampling procedure. The construction of the UCB policy yields another component of regret which is sub-linear.

The first variation we look at is the setting where the change-points are generated uniformly at random. The rest of the setting is maintained as before and the policy has no modifications.

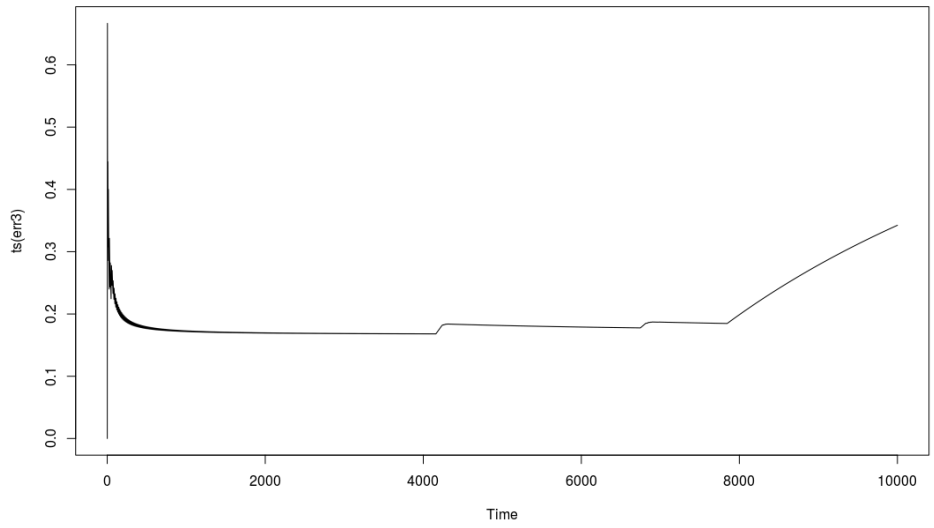

Figure 4.3: Error of the M-UCB policy with random change-points

We can immediately observe that the error for the M-UCB policy with random change-points (i.e. there is no longer the assumption on L) tends to perform very similarly to the standard M-UCB policy when two break-points are far enough from each other as the policy itself has not been changed. However, if two change-points are too close to each other this leads the algorithm to perform poorly and cause a huge increase on the regret. Of course this example cannot generalize the general behaviour of this policy in the modified setting because of the probabilistic nature of the modification we made. To have a better understanding of this setting we are going to discuss the change-points generated at random in the next section.

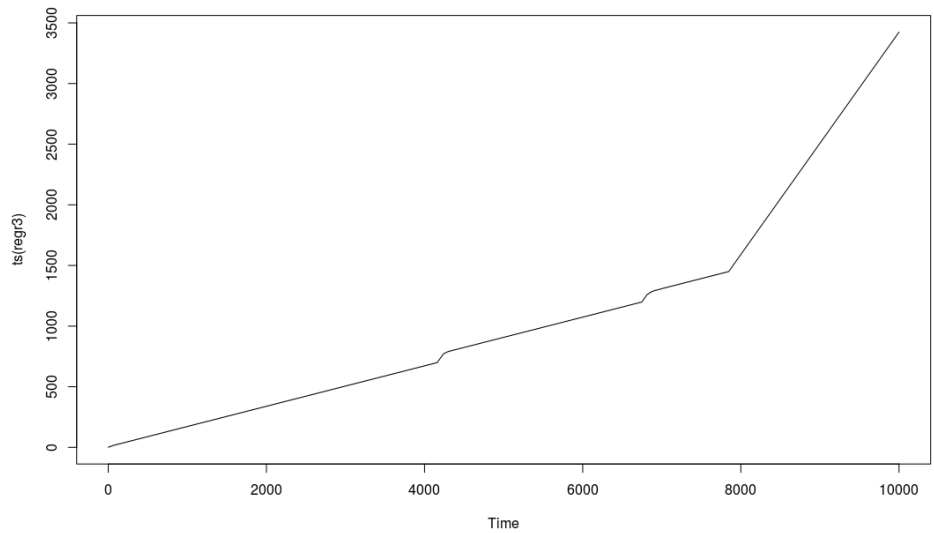

Figure 4.4: Cumulative regret of the M-UCB policy with random change-points

We can see that the slope of the regret line increases significantly if we have two changes close to each other. This is sensible as if the policy does not detect the change than the regret yielded is linear given that it would be like using a standard UCB policy for a non-stationary setting.

Next, we are going to analyze the performance of the algorithm if we only restart the arm where the change was spotted and the assumption of L steps between one change and the next one has to hold on the single arm.
Thus, this looks like a straightforward adaptation of the M-UCB policy for the local changes setting.

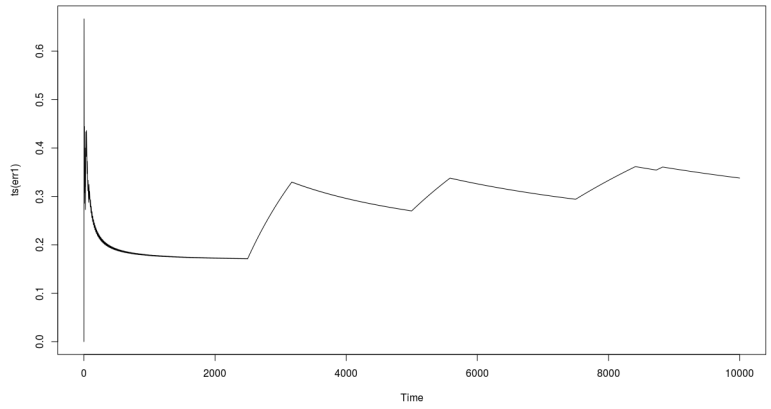

Figure 4.5: Error of the M-UCB policy with single arm restart

This modification implemented does not seems to yield decent performances and it is significantly outperformed by the standard M-UCB. This version of the algorithm seems to struggle especially after each change where the error explodes. This sounds reasonable because of the assumptions made for the M-UCB policy. More specifically, we were assuming that at each breakpoint there was at least a change in distribution with big enough magnitude.
However, if we only restart the arm where the change has been spotted, the other arms that smaller magnitude changes would never have those changes spotted. Therefore, to adapt the assumption to the local changes setting we would have to assume that every change happening on any arm has to have magnitude bigger than some quantity δ.

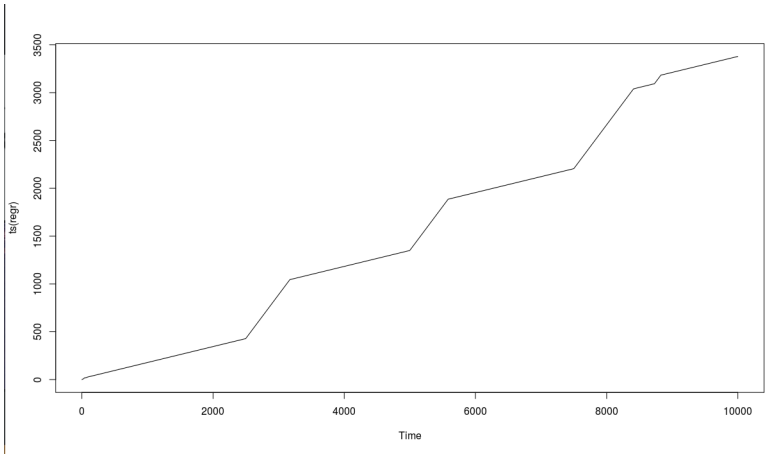

Figure 4.6: Cumulative regret of the M-UCB policy with single arm restart

We want to know whether it is possible to make modifications either to the algorithm either to the assumptions we consider so that the algorithm is able to improve performances in such scenarios.

We start by studying the first modification proposed. Specifically, we show how the distribution of the length of the shortest segment behaves.

4.2 Fixed number of change-points generated at random

Recall that the m points generated at random on the line $[0, 1]$ we denote them with $\theta_1, ..., \theta_m$. Define $D_1, ..., D_{M+1}$ as

$$D_i = \theta_i - \theta_{i-1} \tag{4.1}$$

and $D_{(1)}, ..., D_{(M)}$ are the ordered segments. In the non-stationary bandits context, we can think of these points as global breakpoints. Specifically, let $[0, T]$ be the time interval considered in a specific bandit problem, then $\nu_1 = \lceil T\theta_1 \rceil, ..., \nu_m = \lceil T\theta_m \rceil$ are the change-points. Notice that $S_i = T \cdot D_i$.
Our goal is to study how the length of the shortest segment $D_{(1)}$ is distributed. In fact, if we assume that the change-points are generated at random we can't make assumptions similar to what Cao et al. [2019] did (i.e. $D_i > L$).
Therefore, our aim has to be to find a L such that the probability that L will be greater than that is high and at the same time L is big enough to allow us to perform uniform sampling without making the regret explode. The random fragmentation of the segment $[0, 1]$ has been first studied in Pyke [1965] and then revisited in Pyke et al. [1972].

The first approach we might want to use involves the Cumulative Distribution Function. In fact, if we are given the CDF of $D_{(1)}$ $F_{D_{(1)}}(\cdot)$ then for every $u \in [0, 1]$ there exists $x \in [0, 1]$ such that $F_{D_{(1)}}(x) = P(D_{(1)} \leq x) = u$.
Therefore, if we are given a value u and a minimum number of w we can immediately find the value x so that $F_{D_{(1)}}(x) = u$ and then set $\theta = k\frac{w}{x}$. This means that for $x \leq k \cdot w$, we would have $\theta \geq 1$ which would mean that the regret is maximized.

The second approach we can use requires us to know the theoretical first and second moments (they exist as the probability distribution is defined over a bounded support $[0, 1]$) and therefore to know $\mu_{D_{(1)}}$ the mean and the variance $\sigma_{D_{(1)}}$.
If we are given those values, thanks to Chebyshev inequality, we can write

$$P(|D_{(1)} - \mu_{D_{(1)}}| \geq \sigma_{D_{(1)}} k) \leq \frac{1}{k^2}$$

which basically leads us to write down a confidence interval.
If the CDF of $D_{(1)}$ is given we can derivate the pdf and therefore the expectation and variance.

First of all, we care about the distribution of the shortest interval. This is given to us from Bairamov et al. [2010] (Lemma 1). Specifically,

$$F_{D_{(1)}}(x) = P(D_{(1)} \leq x) = 1 - (1 - x(m + 1))^m.$$

We first compute the PDF:

$$F'_{D_{(1)}}(x) = f_{D_{(1)}}(x) = (m + 1)m(1 - x(m + 1))^{m-1}$$

and then we evaluate the expectation.
We can notice that the segment of minimum length has to be smaller than $\frac{1}{(m+1)}$. To prove this statement, suppose $D_{(1)} > \frac{1}{(m+1)}$. Thus,

$$D_{(1)} > \frac{1}{m + 1} \Rightarrow 1 \geq (m + 1)D_{(1)} > (m + 1) \cdot \frac{1}{m + 1} = 1$$

which leads to a contradiction.
Thus the expectation is evaluated in the interval $[0, \frac{1}{m+1})$.

$$E[D_{(1)}] = \int_0^{1/(m+1)} x(m+1)m(1-x(m+1))^{m-1}dx = (m+1)m\int_0^1 x(1-x(m+1))^{m-1}dx \tag{4.2}$$

$$= (m+1)m\left(\left[\frac{x(1-x(m+1))^m}{m(m+1)}\right]_0^{1/(m+1)} - \int_0^{1/(m+1)}\frac{(1-x(m+1))^m}{m(m+1)}dx\right) \tag{4.3}$$

$$= \left([x(1-x(m+1))^m]_0^{1/(m+1)} - \int_0^{1/(m+1)}(1-x(m+1))^m dx\right) \tag{4.4}$$

$$= \left[x(1-x(m+1))^m - \frac{1}{(m+1)^2}(1-x(m+1))^{m+1}\right]_0^{1/(m+1)} \tag{4.5}$$

$$= \frac{1}{(m+1)^2} \tag{4.6}$$

In a very similar way we evaluate the second moment in order to get the variance of the distribution.

$$E[D_{(1)}^2] = \int_0^{1/(m+1)} x^2(m+1)m(1-x(m+1))^{m-1}dx \tag{4.7}$$

$$= (m+1)m\int_0^{1/(m+1)} x^2(1-x(m+1))^{m-1}dx \tag{4.8}$$

$$\tag{4.9}$$

Let us introduce a change of variable $u = 1 - x(m+1)$ which yields a change of differential $dx = -1/(m+1)du$. Thus,

$$= -\frac{1}{(m+1)^3}\int (u-1)^2(u)^{m-1}du \tag{4.10}$$

$$= -\frac{1}{(m+1)^3}\int u^{m+1} - 2u^m + u^{m-1}du \tag{4.11}$$

$$\tag{4.12}$$

We use linearity tu solve each part of the integral and obtain

$$= -\frac{1}{(m+1)^3}\left(\frac{u^{m+2}}{m+2} - 2\frac{u^{m+1}}{m+1} + \frac{u^m}{m}\right). \tag{4.13}$$

Combining all the terms together and undoing the substitution yields

$$= -\frac{((-m-1)\,x+1)^{m+2}}{(m+1)^3\,(m+2)} + \frac{2\,((-m-1)\,x+1)^{m+1}}{(m+1)^4} - \frac{((-m-1)\,x+1)^m}{m\,(m+1)^3}. \tag{4.14}$$

Multiplying by the $(m+1)m$ leads to

$$= -\frac{m\,((-m-1)\,x+1)^{m+2}}{(m+1)^2\,(m+2)} + \frac{2m\,((-m-1)\,x+1)^{m+1}}{(m+1)^3} - \frac{((-m-1)\,x+1)^m}{(m+1)^2} \tag{4.15}$$

$$= -\frac{(1-(m+1)\,x)^m\left(m\,(m+1)\,x\left((m+1)^2\,x+2\right)+2\right)}{(m+1)^3\,(m+2)}. \tag{4.16}$$

We plug the extremes of the definite integral to obtain

$$= \frac{2}{(m+1)^3(m+2)} \tag{4.17}$$

Thus, the variance is

$$\text{Var}[D_{(1)}] = E[D_{(1)}^2] - (E[D_{(1)}])^2 = \frac{2}{(m+1)^3(m+2)} - \frac{1}{(m+1)^4} \tag{4.18}$$

$$= \frac{m}{(m+1)^4(m+2)} \tag{4.19}$$

Therefore, we can see that the variance tends to be very small.
Thanks to Chebyshev inequality we can write

$$P\left(\left|D_{(1)} - \frac{1}{(m+1)^2}\right| \geq \sqrt{\frac{m}{(m+1)^4(m+2)}} \cdot a\right) \leq \frac{1}{a^2}.$$

This inequality obviously implies that, with probability $1 - \frac{1}{a^2}$ the value lies in the interval

$$\left(\frac{1}{(m+1)^2} - a\sqrt{\frac{m}{(m+1)^4(m+2)}}, \frac{1}{(m+1)^2} + a\sqrt{\frac{m}{(m+1)^4(m+2)}}\right)$$

To make an example, let $m = 2$ (two breakpoints) and $a = 4$ so that with probability 0.9375 the value lies in

$$\left(\frac{1}{9} - 4 \cdot \sqrt{\frac{2}{(3)^4 \cdot (4)}}, \frac{1}{9} + 4 \cdot \sqrt{\frac{2}{(3)^4 \cdot (4)}}\right)$$

$$= \left(\frac{1}{9} - \frac{2\sqrt{2}}{9}, \frac{1}{9} + \frac{2\sqrt{2}}{9}\right)$$

$$= \left(0, \frac{1+2\sqrt{2}}{9}\right)$$

as 0 is the lowerbound on the possible regret.

We investigate here whether using Chebyshev's Inequality for higher moments might lead us to tighter bounds and therefore we compute the third moment and we write the inequality:

$$P(|D_{(1)} - \mu_{D_{(1)}}| \geq \rho a) \leq \frac{1}{a^3}$$

where

$$\rho = E\left[\left(D_{(1)} - \frac{1}{(m+1)^2}\right)^3\right]$$

By computing the third moment we find

$$E\left[\left(D_{(1)} - \frac{1}{(m+1)^2}\right)^3\right] = \frac{6}{(m+1)^4(m^2+5m+6)}$$

which leads to a confidence interval shaped like

$$\left(\frac{1}{(m+1)^2} - a\frac{6}{(m+1)^4(m^2+5m+6)}, \frac{1}{(m+1)^2} + a\frac{6}{(m+1)^4(m^2+5m+6)}\right)$$

For $m = 2$ and $a = 4$, like before, we have

$$\left(\frac{1}{9} - 4\frac{6}{(3)^4(20)}, \frac{1}{9} + 4\frac{6}{(3)^4(20)}\right)$$

$$= \left(\frac{1}{9} - \frac{6}{81 \cdot 5}, \frac{1}{9} + \frac{6}{81 \cdot 5}\right)$$

$$= \left(\frac{1}{9} - \frac{6}{405}, \frac{1}{9} + \frac{6}{405}\right)$$

$$= \left(\frac{1}{9} - \frac{2}{135}, \frac{1}{9} + \frac{2}{135}\right)$$

Which implies that the this interval we found is tighter than the one using the variance.
Furthermore, the interval above is not just tighter. In fact, $s_{(1)}$ is outside of $(\frac{1}{9} - \frac{2}{135}, \frac{1}{9} + \frac{2}{135})$ with probability $0,015$ which gives us a lot of confidence in predicting the size of the minimum segment.
These considerations can be extend directly from $D_{(1)}$ to $S_{(1)}$ as $P(|D_{(1)} - \mu_{D_{(1)}}| \geq \rho a) = P(|S_{(1)} - \mu_{S_{(1)}}| \geq T \cdot \rho a)$ and therefore we just have to multiply L by T. We are now going to analyze the performance of the M-UCB policy when the value of L is tuned as mentioned above. We are also going to revisit the assumption made for the original policy.

Assumption 4.2.1 (Random Change-points)**.** *Let* $L = \frac{1}{(m_T+1)^2} - \frac{24}{(m_T+1)^4(m_T^2+5m_T+6)}$. *The learning agent can choose one between* w *and* γ *s.t.* $P(\nu_{(i+1)} - \nu_{(i)} > L) \approx 0.99, \forall 1 \leq i \leq m_T$, *and b)*$\exists k \leq K \; \forall 1 \leq i \leq m_T$, $\mu^1, ..., \mu^{m_T+1} \in [0,1]^K$ *and*

$$\delta_k^{(i)} \geq 2\sqrt{\frac{\log(2KT^2)}{w}} + 2\sqrt{\frac{\log(2T)}{w}}$$

Before moving into the regret analysis of the algorithm we state some lemmas that will be very useful.
We are going to do it in a similar fashion to Cao et al. [2019]. More specifically, the first two lemmas that deal with the stationary setting are going to be recovered from the section concerning the above mentioned paper as the change-points generated at random, obviously, do not influence the regret behaviour when there are no change-points.

One lemma which yields differences from Cao et al. [2019] version is the one which deals with the probability of achieving a successful detection. In fact, as the length of each segment is random, the probability of detecting it depends as well on how long each segment is.

Lemma 4.2.1 (Probability of achieving a successful detection)**.** *Consider a piecewise-stationary scenario with* $M = 1$ *and recall that* $L = \frac{1}{(m_T+1)^2} - \frac{24}{(m_T^2+5m_T+6)}$ *and that* $P(S_{(1)} \leq TL) \approx 0.01$. *For any* $\mu^1, \mu^2 \in [0,1]^K$ *satisfying* $\delta_{\tilde{k}} \geq 2\frac{b}{2} + c$ *for some* $\tilde{k} \leq K$ *and* $c > 0$, *under Algorithm 1, we have that*

$$P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1\right) \geq \left(1 - 2\exp\left(-\frac{wc^2}{4}\right)\right)\left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right)$$

Proof.

$$
\begin{aligned}
&P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1\right) \\
&= P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1, \theta_2 - \theta_1 > \frac{L}{2}\right) P\left(\theta_2 - \theta_1 > \frac{L}{2}\right) \\
&+ P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1, \theta_2 - \theta_1 \leq \frac{L}{2}\right) P\left(\theta_2 - \theta_1 > \frac{L}{2}\right) \\
&\geq P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1, \theta_2 - \theta_1 > \frac{L}{2}\right) P\left(D_{(1)} > \frac{L}{2}\right) + \\
&+ P\left(\nu_1 < \tau_1 \leq \nu_1 + \frac{TL}{2} \mid \tau_1 > \nu_1, \theta_2 - \theta_1 \leq \frac{L}{2}\right) P\left(\theta_2 - \theta_1 > \frac{L}{2}\right) \\
&\geq P(S_{\tilde{k},w} > b)\left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right) \\
&\geq \left(1 - 2\exp\left(-\frac{wc^2}{4}\right)\right)\left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right)
\end{aligned}
$$

where the first inequality follows from the fact that the probability of any segment being longer than a quantity is for sure bigger than the probability of the shortest segment being grater than that quantity. □

Of course the result makes sense as, given that the shortest segment might be shorter than $\frac{TL}{2}$, the probability that we spot a change is also influenced by the probability that the segment is actually shorter than $TL/2$ influences the probability of a achieving a successful detection.

The last lemma we are going to add discusses the expected delay in the detection. We want to estimate which one is going to be the expected delay in the detection of a change in distribution. Recall that in Cao et al. [2019] was fixed according to the number of samples wanted w and to the uniform sampling coefficient θ set as a function of w, m_T, K and T. Therefore the problem and the player parameters were defining how long the segment had to be.
In the case we want to address now, the length of the interval is given by the probabilistic setting we work in. In fact, L is equal to $L = \frac{1}{(M+1)^2} - \frac{24}{(m_T+1)^4(m_T^2+5m_T+6)}$, so one thing we might want to do is, given w, to set γ so that $L = w\left\lceil\frac{K}{\gamma}\right\rceil$.
Thus,

$$
\begin{aligned}
L &= \frac{1}{(m_T+1)^2} - \frac{24}{(m_T+1)^4(m_T^2+5m_T+6)} \\
\gamma &= w\frac{K}{L} = w\left(\frac{K}{\frac{1}{(m_T+1)^2} - \frac{24}{(m_T+1)^4(m_T^2+5m_T+6)}}\right)
\end{aligned}
$$

Given this, the lemma to deal with the expected detection delay is again the same as Lemma 7 as, given that the change in distribution has been spotted, the delay in spotting it does not depend on the randomness of the change-points. By comparing these lemmas to Cao et al. [2019] (which are very similar in most aspects) we can notice that the main difference is, as always, in the different definition given of L.
We now proceed to the actual statement of the theorem and its proof.

Theorem 4.2.1. *Running Algorithm 1 with appropriate choices of w and γ satisfying Assumption*

2 and $b = \sqrt{w \log(2KT^2)/2}$*, we have*

$$\mathcal{R}(T) \leq \sum_{i=1}^{m_T} \tilde{C}_i + \gamma T + \sum_{i=1}^{m_T} \min\left(\frac{TL}{2}, (\left\lceil b/\Delta_{\tilde{k}}^{(i)} \right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil\right) + m_T \left(F_{S_{(1)}}\left(\frac{TL}{2}\right) \cdot T + 2\left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right) + \frac{1}{T}\left(F_{S_{(1)}}\left(\frac{TL}{2}\right) - 1\right)\right)$$

Proof. Recall that $L = \frac{1}{(m_T+1)^2} - \frac{24}{(m_T+1)^4(m_T^2+5m_T+6)}$ and while one between w and θ is chosen by the player the other is set such that $L = w\left\lceil \frac{K}{\gamma} \right\rceil$.

According to our setting $P(D_{(1)} - \frac{1}{(m_T+1)^2} < \frac{24}{(m_T+1)^4(m_T^2+5m_T+6))}) \approx 0.01$. Define events $F_i = \{\tau_i > \lceil T\theta_i \rceil\}$, $W_i = \{\tau_i < \lceil T\theta_i \rceil + \frac{TL}{2}\}$ and $G_i = \{\lceil T\theta_{i+1} \rceil - \lceil T\theta_i \rceil > L/2\}, i \in \{1, ..., m_T\}$. It follows that the event $F_i \cap W_i \cap G_i$ is the event where the $i+1$-th change is not too close to the i-th one and it can be detected correctly and efficiently.
Define $R(T) = \sum_{t=1}^{T} \max_{k \in K} X_{k,t} - X_{A_t,t}$ and thus $\mathcal{R}(T) = E[R(T)]$. Thus

$$\begin{aligned}
\mathcal{R}(T) = E[R(T)] &\leq E[R(T)\mathbb{1}\{F_1\}] + E[R(T)\mathbb{1}\{\overline{F}_1\}] \\
&\leq E[R(T)\mathbb{1}\{F_1\}] + T \cdot (1 - P(\mathcal{F}_i)) \\
&\leq E[R(T)\mathbb{1}\{F_1\}] + 1 \\
&\leq E[R(\nu_1)\mathbb{1}\{F_1\}] + E[R(T) - R(\nu_1)] + 1 \\
&\leq \tilde{C}_1 + \gamma\lceil T\theta_i \rceil + E[R(T) - R(\lceil T\theta_i \rceil)] + 1
\end{aligned}$$

where the second inequality follows from the classic upper-bounding, the third one follows from Lemma 3.4.2 by setting $b = \sqrt{w \log(2KT^2)/2}$ and the last one follows from Lemma 3.4.1 (the segment of time before ν_1, time of the first change, is clearly stationary). Therefore, we have found an upper-bound on the first segment of time.
We have now to upper-bound $E[R(T) - R(\lceil T\theta_i \rceil)]$.

$$\begin{aligned}
E[R(T) - R(\lceil T\theta_i \rceil)] &\leq E[R(T) - R(\nu_1) \mid F_1, W_1] + T \cdot (1 - P(F_1, W_1)) \\
&\leq E[R(T) - R(\nu_1) \mid F_1, W_1] + T \cdot (1 - P(W_1 \mid F_1)P(F_1)) \\
&\leq E[R(T) - R(\lceil T\theta_i \rceil) \mid F_1, W_1] + T \cdot \left(1 - \left(1 - \frac{1}{T}\right)^2 \left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right)\right) \\
&\leq E[R(T) - R(\nu_1) \mid F_1, W_1] + 2 - \frac{1}{T} + F_{S_{(1)}}\left(\frac{TL}{2}\right) \cdot T - 2F_{S_{(1)}}\left(\frac{TL}{2}\right) \\
&+ \frac{1}{T}F_{S_{(1)}}\left(\frac{TL}{2}\right) \\
&\leq E[R(T) - R(\nu_1) \mid F_1, W_1] + F_{S_{(1)}}\left(\frac{TL}{2}\right) \cdot T + 2\left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right) \\
&+ \frac{1}{T}\left(F_{S_{(1)}}\left(\frac{TL}{2}\right) - 1\right)
\end{aligned}$$

where the third inequality follows from Lemma 4.2.1. The fourth and the fifth inequality follow from further simplifications of the expression.
To make things more clear, the term $\left(1 - F_{S_{(1)}}\left(\frac{L}{2}\right)\right)$ is an upperbound on $P(G_1)$ as showed in the proof of Lemma 4.2.1.
We can notice that having a distribution over the change-points influences the structure of the regret. In fact, it yields a term which grows linearly with T. How big this term will be depends only on the number of change-points, as it is the only parameter that define the above mentioned CDF.

Now, we only have to bound the term of the regret conditioned on the "good events" to happen.

$$\begin{aligned}E[R(T) - R(\nu_1)|F_1, W_1] &\leq E[R(T) - R(\tau_1)|F_1, W_1] + E[R(\tau_1) - R(\nu_1)|F_1, W_1] \\ &\leq \tilde{E}[R(T - \nu_1)] + E[\tau_1 - \nu_1|F_1, W_1] \\ &\leq \tilde{E}[R(T - \nu_1)] + \min\left(\frac{TL}{2}, (\left\lceil b/\Delta_{\tilde{k}}^{(1)}\right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil\right) / \left(1 - \frac{1}{T}\right)\end{aligned}$$

where $\tilde{E}$ is the expectation according to the bandit problem starting from the second segment. The second inequality follows from the renewal property of the algorithm (after spotting a change everything restarts) whereas the third comes from Lemma 3.4.4.
Thus,

$$\begin{aligned}E[R(T)] \leq \tilde{E}[R(T - \nu_1)] + \tilde{C}_1 + \gamma\nu_1 + F_{S_{(1)}}\left(\frac{TL}{2}\right) \cdot T + 2\left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right) \\ + \frac{1}{T}\left(F_{S_{(1)}}\left(\frac{TL}{2}\right) - 1\right) + \min\left(\frac{TL}{2}, (\left\lceil b/\Delta_{\tilde{k}}^{(1)}\right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil\right)\end{aligned}$$

Thanks to the recursion we defined, we can conclude that

$$\begin{aligned}E[R(T)] \leq \sum_{i=1}^{m_T} \tilde{C}_i + \gamma T + \sum_{i=1}^{m_T} \min\left(\frac{TL}{2}, (\left\lceil b/\Delta_{\tilde{k}}^{(i)}\right\rceil + 3\sqrt{w}) \cdot \lceil K/\gamma \rceil\right) \\ + m_T\left(F_{S_{(1)}}\left(\frac{TL}{2}\right) \cdot T + 2\left(1 - F_{S_{(1)}}\left(\frac{TL}{2}\right)\right) + \frac{1}{T}\left(F_{S_{(1)}}\left(\frac{TL}{2}\right) - 1\right)\right)\end{aligned}$$

□

Before moving on, we show a simulation based on the algorithm proposed above. Let the problem be similar to the one we defined in the simulations section previously. Let the problem have three arms and an observed time horizon of 10000.

The three arms have means switching as mentioned previously: arm 1 passes from mean 0.9 to mean 0.1 after each change, arm 2 does the opposite whereas arm 3 remains constant at 0.5. We allow the points to be generated at random and we obtain changes generated at random at times 5844, 7207, 9631. This setting seems to be interesting as, while the first and last changes are far apart enough, the second is quite close to the first one and would probably cause problems if we used the normal M-UCB.

The error itself is going to be higher than the one we perceive with the regular M-UCB before the anomaly, but will prevent it from exploding after the changes close to each other.

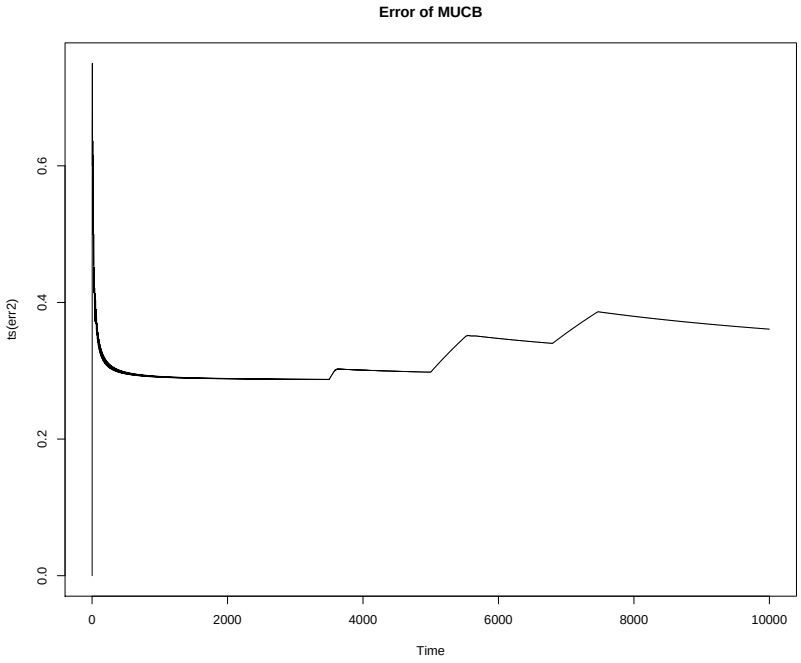

Figure 4.7: Error of the M-UCB policy with change-points generated at random

The plot highlights the fact that the algorithm, at the cost of a higher exploration price, is able to detect all the three changes happening even if they are close in a way that with the standard setting would not be detected. Therefore, this setting looks stable in case someone wanted to use M-UCB where changes can happen randomly close to each other, however it will require an higher cost in exploration. However, someone might argue that, again, known the number of change-points, performing detection might not be optimal. Furthermore, the D-UCB algorithm would be able to perform on a setting where the changes are generated at random without any probability of incurring in a linear regret because of a missdetection. Moreover, the detector seems to be an useful tool when we have to deal with an unknown number of changes. Therefore, we introduce another scenario. We assume to have a distribution over the possible number of change-points without knowing the exact number.

4.3 Extension: Random number of change-points

Suppose that the change-points are generated according to some probability distribution but here we suppose that the number of change-points M is described from a random variable as well. We will call the set containing the possible values of M as $\mathcal{M}$. We will, of course, only consider finite sets for $\mathcal{M}$.
Let $\mathcal{M} = \{M', M'', M'''\}$ where the maximum element of this set is still much smaller than N. Let the prior probability that we have M' change-points be denoted by $p^0_{M'} = P(M = M')$.
Of course it holds that

$$\sum_{m \in \mathcal{M}} p^0_m = 1.$$

Here we are considering that we may have some prior knowledge on the number of change-points. If we only know the possible numbers of change-points we can set the prior distribution to be non-informative so that the initial probability values don't have an heavy influence on the posterior ones.

The goal is to update this probabilities when new information on the process generating the segment becomes available (i.e. after change-point).
However, it might be interesting to update it also when a change does not happen for a long enough time, this would prevent maybe the problem that if a lot of changes happen in the early stages and there is none for a long time the estimates would remain biased for a very long period.
Two possible approaches concerning the ways we can tune the number of change-points parameter M in the algorithm follow below.
The first approach would require to set the value of M to the one $\in \mathcal{M}$ which maximizes the posterior probability of having generated such segments.
The second one is to set M to a mixture of the possible values it can assume where the parameters of the mixture are given by the posterior probabilities of each one of those or a by a function of them.

Clearly the two approaches have different pros and cons: the first one seems to be more risky as, if we are able to spot the correct value in $\mathcal{M}$, we would have an unbiased estimate but in case the early events don't allow us to spot the right value than we would probably incur a lot of regret. The second approach is safer as it brings in some bias but it does not go all in on an option which might be wrong.
In order to determine the posterior probabilities, however, we have to give the probability distribution functions of the sorted segments. Recall that if we have M change-points, we have $M+1$ segments and we indicate by D_i the random variable of the i-th segment.
Then, if we have let M be the number of change-points, the distribution of each segment is Pyke [1965]

$$F_{D_i}^M(x) = F_{D_1}(x) = 1 - (1-x)^M$$

whereas for any couple of segments $(D_i, D_j),\ j \neq i$, is

$$\begin{aligned} F_{(D_i,D_j)}^M(x,y) = F_{(D_1,D_2)} &= P(U_{(1)} \leq x, U_{(2)} - U_{(1)} \leq y) \\ &= M \int_0^x \left\{ 1 - \left(1 - \frac{y}{1-u} \right)^{M-1} \right\} (1-u)^{M-1} du \\ &= 1 - \{(1-x)^M + (1-y)^M - (1-x-y)^M\}. \end{aligned}$$

Equivalently, the corresponding density functions are

$$\begin{aligned} f_{D_i}^M(x) &= M(1-x)^{M-1} \\ f_{(D_i,D_j)}^M(x,y) &= M(M-1)(1-x-y)^{M-2}. \end{aligned}$$

We will need also higher dimensional joint distributions which I still have to find.
Given this, let's start considering the mixture approach.

Recall that for each $m \in \mathcal{M}$ we define the prior probability as p_m^0. Thus, we start the game by setting

$$M_0 = \sum_{m \in \mathcal{M}} p_m^0 \cdot m$$

or, alternatively

$$M_0 = \sum_{m \in \mathcal{M}} g(p_m^0) \cdot m$$

where $g : [0,1] \to [0,1]$ is a function such that

$$\sum_{m \in \mathcal{M}} g(p_m^0) = 1.$$

We keep playing until we spot a change after l steps. After that we update the parameters of the mixture.
We first of all compute the posterior probability

$$P(M = M' | D_1 = l) \propto f_{D_1}^{M'}(l) \cdot p_{M'}^0$$

and thus we can write

$$p_M^1 = P(M = M' | D_1 = l) = \frac{f_{D_1}^{M'}(l) \cdot p_{M'}^0}{\sum_{m \in \mathcal{M}} f_{D_1}^m(l) \cdot p_m^0}.$$

This would lead us to update the mixture parameters to

$$M_1 = \sum_{m \in \mathcal{M}} p_M^1 \cdot m$$

or equivalently to

$$M_1 = \sum_{m \in \mathcal{M}} g(p_M^1) \cdot m.$$

We then keep playing until the second change $D_2 = l'$ happens and we then write, in a similar fashion,

$$P(M = M' | D_1 = l, D_2 = l') \propto f_{(D_1, D_2)}^{M'}(l, l') \cdot p_{M'}^0$$

and thus,

$$p_M^2 = P(M = M' | D_1 = l, D_2 = l') = \frac{f_{(D_1, D_2)}^{M'}(l, l') \cdot p_{M'}^0}{\sum_{m \in \mathcal{M}} f_{(D_1, D_2)}^m(l, l') \cdot p_m^0}.$$

which leads to a further parameter update.
The idea is to keep doing this, of course if we spot m changes, with some elements $\in M$ smaller than m, we will drop those elements from the mixture.

If we decide to generate a fixed amount of change-points m_n in the interval $[0, n]$ and we then consider the time-line $[0, T]$ (with $n \geq T$), the number of change-points we are going to have is random. Thus, as the number of change-points $m_T \leq m_n$ is random, we can define a probability distribution on the number of change-points at time T and use these probabilities as the prior parameters for the mixture with Bayesian updating proposed above.

We show below an example of functioning of this algorithm. Consider a scenario where three changes happen at $2500,\ 6000$ and 8500. The means of each arm remain the same as the previous settings, arm 1 swaps from mean reward 0.9 to 0.1, arm 2 the opposite and arm 3 remains at constant mean 0.5. We are going to assume we don't know the number of changes and that we know that the possible values are $2,\ 3,\ 4$. We also assume to have a uniform prior distribution on the values so that it does not influence the final choice made by the algorithm. The value of θ is set according to the estimate of the number of change-points. The plot of the behaviour of the error is reported below.

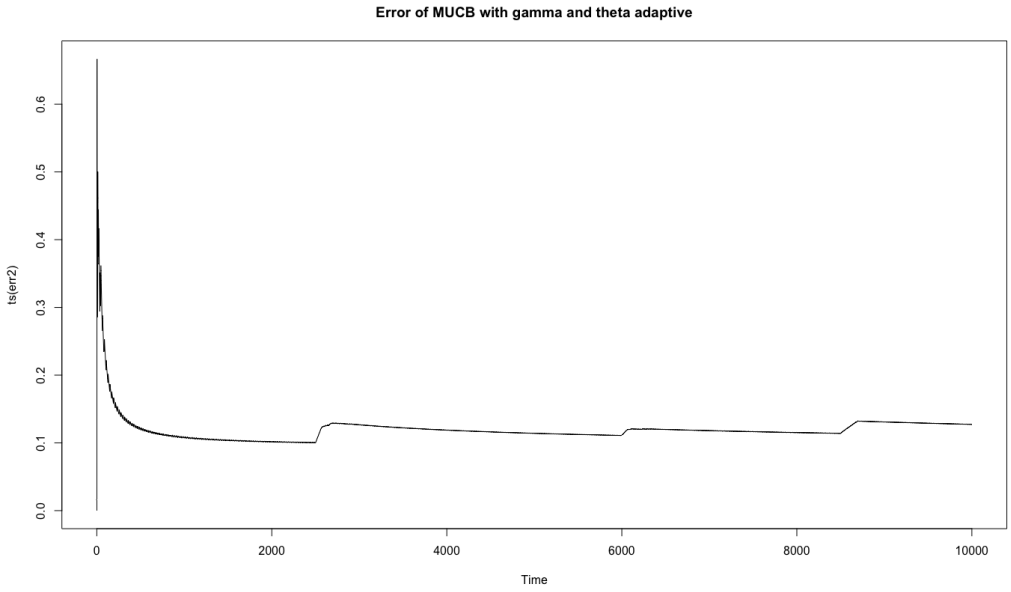

Figure 4.8: Cumulative regret of the M-UCB policy in its standard setting

The estimate of the posterior probabilities are 0.00 for the value 2 (as obvious, as it spotted three changes), 0.636 for the value 3, the correct one, and 0.364 for 4, as it still considers that there might be more changes coming in the remainder of the line. This idea would give a sensible explanation to the use of a change-point detector in a bandit algorithm as, without knowing the exact number of changes, we can aim to set the parameters to their right value only by trying to spot them. Therefore the goals for new works might be to implement new strategies that take in account a distribution over the possible number of changes. Consider that the change-point detector might be a more refined algorithm like a CUSUM algorithm or a Likelihood-Ratio test type of detector like Besson and Kaufmann [2019].

Chapter 5

Conclusions

In this work we have provided an analysis of the Non-Stationary Bandit setting. To achieve a complete overview we analysed policies for the stationary setting and for the adversarial one as well. These were useful to introduce the reader to the bandit problem and to investigate the concept behind the UCB type of policies.

In Chapter 2 we gave a definition of the problem from a generic point of view before moving into distinguishing between the stochastic setting and the adversarial one. After a brief introduction for both settings we put our focus mostly on the stochastic one. We defined the UCB type of policies and we showed how and why they work. Following this, we introduce the UCB1 strategy (Auer et al. [2002a]), the (α, ϕ)-UCB one, the KLUCB one (Garivier and Cappé [2011]) and the ϵ-greedy policy that does not belong to the UCB family. We also showed a lower bound on the expected regret for the general stochastic setting in order to know what the best achievable performance is. We concluded this chapter by discussing the adversarial setting and by providing an analysis of the Exp3 strategy (Bubeck et al. [2012]) for it.

In Chapter 3 we provided an analysis of the Non-Stationary setting and of some of the most used policies for said setting. More specifically, we start the chapter by introducing some new notation and basic definitions and notions. We then introduce two policies by Garivier and Moulines [2011], namely D-UCB and SW-UCB, showing the expect regret bound they manage to achieve. We also provided a lower bound for the expected regret that any policy can achieve in a piecewise stationary type of problem. Chapter 3 is concluded with the presentation of the M-UCB policy, introduced in the literature by Cao et al. [2019]. This algorithm presents as main difference with the others the presence of a change-point detector to check for changes in distribution. We showed the regret analysis of this policy to verify whether introducing this component had a strong impact on the expected regret, concluding that the change-point detection as presented in the M-UCB policy does not look to improve the performances yieled by the other algorithms presented for this setting and that the change-point detector might be more useful in a situation where the decision maker does not have precise knowledge about the number of changes occurring in the game.

Chapter 4 concludes this work and its aim was to come up with possible ideas for new algorithms. We started the chapter with some quick simulation useful to understand which idea might be more promising and why. The simulations were operated on variations of the M-UCB policy using the same algorithm in slightly different environments. For example, the first variation considered was M-UCB where the break-points are not previously fixed but are generated uniformly at random on the time segment. The simulation provided insight on how to modify the settings of the policy to make it work in such an environment. We therefore proposed a different setting of the parameters that would guarantee similar performances to the ones achieved by M-UCB in its regular setting. To do so we introduced some concepts coming from the Uniform Spacings liter-

ature (Pyke [1965], Pyke et al. [1972]) which allow us to determine the probability distribution of the length of each segment. After performing a regret analysis over this reformulated problem we introduced another variation that might make the change-point detection more suitable for the problem. We can suppose that the number of change-points is not known in advance but not completely unknown as well . More specifically, let the number of change-points be one of finitely many options. Then we propose as future idea a variation based on the representation of the number of change-points as a mixture of the possible one with the parameters of the mixture constantly updated through Bayesian updating.

Bibliography

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Ismihan Bairamov, Alexandre Berred, and Alexei Stepanov. Limit results for ordered uniform spacings. *Statistical Papers*, 51(1):227, 2010.

Lilian Besson and Emilie Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*, 2019.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Yang Cao, W Zheng, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure for piecewise-stationary bandit: a change-point detection approach. *AISTATS,(Okinawa, Japan)*, 2019.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

Baskiotis Teytaud Sebag Hartland, Gelly. Multi-armed ban- dit, dynamic environments and meta-bandits. *nIPS-2006*, 2006.

Szepesvari S. Kocsis L. Discounted ucb. In *In 2nd PASCAL Challenges Workshop, Venice, Italy*, 2006.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

E. S. PAGE. CONTINUOUS INSPECTION SCHEMES. *Biometrika*, 41(1-2):100–115, 06 1954. ISSN 0006-3444. doi: 10.1093/biomet/41.1-2.100. URL https://doi.org/10.1093/biomet/41.1-2.100.

Ronald Pyke. Spacings. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27 (3):395–436, 1965.

Ronald Pyke et al. Spacings revisited. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*. The Regents of the University of California, 1972.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.

Alan Willsky and H Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic control*, 21(1):108–112, 1976.